AN INVESTIGATION OF THE EFFECTS OF CORRELATION, AUTOCORRELATION, AND SAMPLE SIZE IN CLASSIFIER FUSION

THESIS

Nathan J. Leap, Captain, USAF

AFIT/GOR/ENS/04-06

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/GOR/ENS/04-06

AN INVESTIGATION OF THE EFFECTS OF CORRELATION, AUTOCORRELATION, AND SAMPLE SIZE IN CLASSIFIER FUSION

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Engineering and Environmental Management

Nathan J. Leap, BS

Captain, USAF

March 2004

AFIT/GOR/ENS/04-06

AN INVESTIGATION OF THE EFFECTS OF CORRELATION, AUTOCORRELATION, AND SAMPLE SIZE IN CLASSIFIER FUSION

Nathan J. Leap, BS
Captain, USAF

Approved:

//signed//	03 Mar 04
Kenneth W. Bauer (Chairman)	date
//signed//	03 Mar 04
Mark E. Oxley (Member)	date

AFIT/GOR/ENS/04-06

Abstract

This thesis extends the research found in Storm, Bauer, and Oxley, 2003. Data correlation effects and sample size effects on three classifier fusion techniques and one data fusion technique were investigated. Identification System Operating Characteristic Fusion (Haspert, 2000), the Receiver Operating Characteristic "Within" Fusion method (Oxley and Bauer, 2002), and a Probabilistic Neural Network were the three classifier fusion techniques; a Generalized Regression Neural Network was the data fusion technique. Correlation was injected into the data set both within a feature set (autocorrelation) and across feature sets for a variety of classification problems, and sample size was varied throughout. Total Probability of Misclassification (TPM) was calculated for some problems to show the effect of correlation on TPM. Feature selection was performed in some experiments to show the effects of selecting only certain features. Finally, experiments were designed and analyzed using analysis of variance to identify what factors had the most significant impact on fusion algorithm performance.

Acknowledgments

I would like to express my sincere thanks and appreciation to my faculty advisor, Dr Kenneth Bauer, for his support and guidance throughout the thesis process. His expertise and insights were invaluable, and he made the thesis effort an enjoyable experience. His ability to develop ingenious solutions to difficult problems taught me valuable lessons that not only helped me through the thesis effort but that I will use throughout my career. I would also like to thank my reader, Dr Mark Oxley, for his assistance with MatLab code and mathematical notation.

I would also like to thank my classmates for their consistent support throughout the past 18 months. Thank you for all of your help at AFIT. Finally and most importantly, I would like to thank my wife for her constant love and support; I could not have done it without you.

Nathan J. Leap

Table of Contents

List of Figures

List of Tables

AN INVESTIGATION OF THE EFFECTS OF CORRELATION, AUTOCORRELATION, AND SAMPLE SIZE IN CLASSIFIER FUSION

I. Introduction

Background

In general, a classification problem is a situation where it is of interest to describe members of a specific number of classes by certain attributes, or features, that the members possess. In the Air Force, a common classification problem is trying to classify targets as hostile, friendly, neutral, or otherwise, based upon certain features that each class possess. In Air Force Doctrine, the Air Force warns its members not to strike targets based on single source intelligence; at some level, intelligence information should be fused together (AFPAM 14-210). Many fusion models are based on the assumption that each of the inputs, in some cases individual classifiers, to the model is independent. In the real world, there are times when classifiers are looking at similar information and are not actually independent; that is, knowing the output of one classifier provides information about the output of another classifier. The more dependent one classifier is on the other, the less new information is present from the additional classifier. In addition, if a classifier is observing a target through time, each observation that it takes may not be independent of the previous observations. Again, this means that less new information is present if the observations are correlated in time. Not much is known about the performance of fusion techniques when faced with correlation (Willett, et al, 2000). In addition, the number of observations that are gathered can significantly impact the performance of an individual classifier and thus fusion of individual classifiers. If

there are many features present in each observation, it may be beneficial to only select certain features that provide more information than others. This thesis examines the effects of sample size, both of these types of correlation, and feature selection on four different fusion models in a variety of different problems. Four different fusion models are used throughout this thesis. Two of these models assume that each classifier is independent from the other classifier, Identification System Operating Characteristic (ISOC) (Haspert, 2002) and Receiver Operating Characteristic (ROC) "Within" (Oxley and Bauer, 2002); two of these models make no such assumption, Probabilistic Neural Network (PNN) and One Big Network (OBN).

Problem Statement

In this thesis, the effects of sample size, two types of correlation, and feature selection on four different fusion models in a variety of different problems are examined. Each problem is constrained to a two-class problem where the two classes are friendly and hostile, and for each problem, only two classifiers are fused via each fusion method. The fusion models are first tested on simple problems, and the problems increase in degree of complexity.

Outline of Thesis

This thesis is divided into five chapters: Introduction, Literature Review, Methodology, Findings and Analysis, and Conclusions. The following is a brief description of the contents of each chapter.

Chapter 1: Introduction – This chapter discusses the background, problem statement, and outline of the thesis.

Chapter 2: Literature Review – This chapter summarizes the pertinent literature on reasons for fusing classifiers, four types of classifier fusion, statistical independence of classifiers, and sample size considerations.

Chapter 3: Methodology – This chapter describes the general methodology employed in this thesis. It describes the two types of correlation, the data generation process for each of the different problems, application issues for each of the four fusion methods, feature selection, TPM, and sample size variation.

Chapter 4: Findings and Analysis – This chapter describes the findings and analysis for each of the problems explored in this thesis.

Chapter 5: Conclusion – This chapter summarizes the results of the research and provides suggestions for future research.

II. Literature Review

Introduction

This chapter provides a summary of the pertinent literature available on reasons for fusing classifiers as well as classifier fusion techniques. First, the Air Force mandates that fusion take place when attacking a target; reasons for fusing classifiers are given in Air Force Doctrine. Next, the statistical independence of the classifiers assumption is discussed. Then, details from each of the four fusion models are provided. Finally, some sample size considerations are discussed.

Air Force Targeting

Any time the United States Air Force prepares to attack a target, there are six steps necessary for the mission: detection, location, combat identification, decision, execution, and assessment (AFPAM 14-210, 1998). Often, combat identification is perceived as the weakest of these six steps since no sensor performs perfectly all of the time (Haspert, 2000). Commanders should be cautious of even the best intelligence on a target, especially when it comes from a single source (AFDD 2-1, 2000). Normally, intelligence on a target should not be based on a single source (AFPAM 14-210, 1998). This leads to the implementation of multiple sensors; combining information from multiple sources, data fusion, increases the confidence in the combat identification step (AFPAM 14-210, 1998). Also, data fusion increases the reliability of the information and makes it more credible and reliable (AFPAM 14-210, 1998). Combining outputs from multiple sensors in order to get a better overall classification accuracy is called sensor fusion. This thesis focuses on improving combat identification through sensor fusion.

Statistical Independence

Many sensor fusion methods make the assumption that the individual sensors are statistically independent. If two or more sensors are statistically independent, it makes sense to combine these sensors to make a better overall decision. However, if two or more sensors are identical, no more information can be gained by adding the additional sensors (Shipp and Kuncheva, 2002). In real world, there are times when the features observed by one sensor are correlated with features observed by another sensor; this creates statistically dependent sensors. Little is known about how sensor fusion methods perform in the presence of statistical dependence since most methods assume statistical independence (Willett, et al, 2000). In previous research, the Gaussian shift-in-means problem was examined in the presence of correlation, and this problem can be broken down into three regions: the "good," the "bad," and the "ugly (Willet, et al., 2000)." It was shown that for the logical "and" and logical "or" rules, any problem in the "good" threshold region should use optimal sensor rules just like those used in the presence of statistical independence (Willet, et al., 2000).

Fusion Methods

Three methods of sensor fusion are used in this thesis: Identification System Operating Characteristic (ISOC) Fusion, Receiver Operating Characteristic (ROC) "Within" Fusion, and Probabilistic Neural Network (PNN) Fusion. Although these methods take different approaches, they have the same overall goal. Each sensor fusion method seeks to improve upon the classification accuracy of a single sensor by combining the outputs of multiple sensors into a single output. Figure 1 shows the overall sensor fusion process.

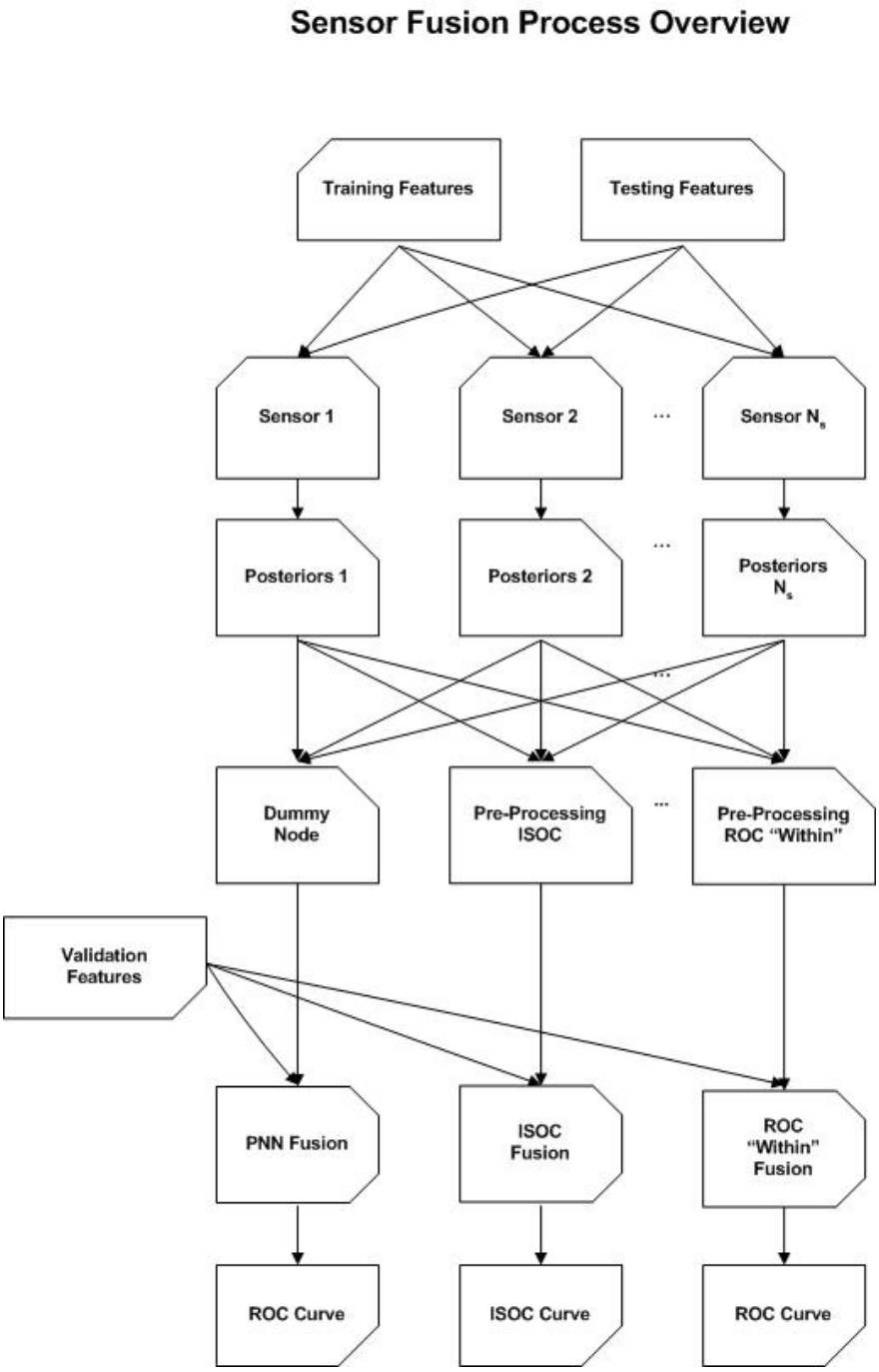


Figure 1: Sensor Fusion Process Overview.

ISOC Fusion Method

The Identification System Operating Characteristic (ISOC) method determines the optimal fusion rule set for a given threshold through a novel algorithm (Haspert, 2000).

This is a paradigm shift from the traditional fixed rules. Although fixed rules are easy to employ, they are not usually optimal (Haspert, 2000). On the other hand, adaptive rules such as Bayesian techniques find an optimal ID sensor fusion rule based on data from a specific target instead of fixing a rule across all data sets (Haspert, 2000). These adaptive rules are based on the results of individual classifiers through a sensor probability matrix (Haspert, 2000).

Sensor Probability Matrices

Combat Identification Systems (CID) take inputs from individual sensors and combine these inputs to form an overall classification (Haspert, 2000). The output of each individual sensor for a given threshold can be output in the following format shown in Table 1.

Table 1: Sensor Probability Matrix.

		Indication	
Truth		"H"	"F"
	H	P("H"|H)	P("F"|H)
	F	P("H"|F)	P("F"|F)

The values in this table will change as the threshold changes for each individual sensor. The rows of this matrix represent the possible types of targets that each individual sensor can observe where F represents friend and H represents hostile, and the columns of this matrix represent the possible sensor outputs. P("H"|H) is the conditional probability of the sensor designating the target as "H" given the target is a hostile. The other conditional probabilities are similar. In this case, the indication "H" is considered a positive and the indication "F" is considered a negative. Therefore, P("H"|H) is the

probability of true positive, P("H"|F) is a false positive, P("F"|H) is a false negative, and P("F"|F) is a true negative.

Combat ID System States

Let N_s be the number of sensors on a target. Let i denote the index of those sensors where $1 \leq i \leq N_s$. Let n_i denote the number of indicator states for sensor i. Let k_i be a specific output state for sensor i. Using these definitions, there will be N total distinct configurations of the Combat ID system where

$$N = \prod_{i=1}^{N_s} n_i \text{ (Ralson, 1998)}$$

Let $S = \bigcup_{j=1}^{N} S_j$ be all possible configurations of the CIS where S_j is the j^{th} output state of the CIS and $1 \leq j \leq N$. Each $S_j = \{s_1^j, s_2^j, s_3^j, ..., s_{N_s}^j\}$ where s_i^j is the state of the i^{th} sensor in the j^{th} configuration (Storm, Bauer, and Oxley, 2003). Thus, *S* is an N x N_s matrix. Table 2 shows the possible combinations of *S*.

Table 2: Sensor Output State Combinations.

j	S_j
1	$(s_1^1, s_2^1, s_3^1, ..., s_{N_s}^1)$
2	$(s_1^2, s_2^2, s_3^2, ..., s_{N_s}^2)$
3	$(s_1^3, s_2^3, s_3^3, ..., s_{N_s}^3)$
. .	. .
. .	. .
N	$(s_1^N, s_2^N, s_3^N, ..., s_{N_s}^N)$

For the two sensor, two state case, S is a 4 x 2 matrix. Table 3 shows the possible combinations of S for this case.

Table 3: Sensor Output State Combinations, Two Sensors and Two Output States.

J	S_j
1	(s_1^1, s_2^1) =("H","H")
2	(s_1^2, s_2^2) =("H","F")
3	(s_1^3, s_2^3) =("F","H")
4	(s_1^4, s_2^4) =("F","F")

Under the assumption that all sensors are independent, the probability of a sensor configuration given truth simply equals the multiplication of the probabilities of the individual sensors in that configuration given truth (Ralston, 1998). This is given by the following equation

$$P(S_j \mid T) = \prod_{i=1}^{N_s} P(s_i^j \mid T).$$

For the two-class problem, in the previous equation, $T \in \{H,F\}$. For each possible output combination, S_j, the probabilities $P(S_j|H)$ and $P(S_j|F)$ must be calculated. Since every potential target, regardless of whether it is friendly or hostile, will put the CIS into some state, $\sum_{j=1}^{N} P(S_j \mid F) = \sum_{j=1}^{N} P(S_j \mid H) = 1$ (Ralston, 1998). After all these probabilities have been calculated, the fusion rules must be defined (Ralston, 1998).

Fusion Rules

There will be times when the CIS will receive conflicting indications from the individual sensors. The fusion rules resolve all of these conflicts by specifying when to

declare hostile and when not to declare hostile (Ralston, 1998). In the two state problem, a complete ID fusion rule can be expressed as an N-dimensional vector R = (r_1, r_2, … , r_N) where $r_j \in \{0,1\}$ and j = 1, 2,…, N (Ralston, 1998). In this case, each element of R corresponds to an element of S. If $r_j = 1$, rule S_j should be included in the rule set (Ralston, 1998). For example, in the two-class, two sensor problem defined above, if R = (1, 0, 1, 0), rules S_1 = ("H","H") and S_3 = ("F","H") should be included in the rule set. Thus, a target will be declared hostile if either S_1 or S_3 occurs. For each specific fusion rule, the probability of that rule given truth can be found with the following equation.

$$P(R \mid T) = \sum_{j=1}^{N} P(S_j \mid T) \cdot r_j \ .$$

By substituting the equation above,

$$P(R \mid T) = \sum_{j=1}^{N} (\prod_{i=1}^{N_s} P(s_i^j \mid T)) \cdot r_j \ .$$

For the two-class problem where $T \in \{H,F\}$, an equation for each element of T follow:

$$P(R \mid H) = \sum_{j=1}^{N} (\prod_{i=1}^{N_s} P(s_i^j \mid H)) \cdot r_j$$

$$P(R \mid F) = \sum_{j=1}^{N} (\prod_{i=1}^{N_s} P(s_i^j \mid F)) \cdot r_j$$

Now that these probabilities have been calculated, R must be chosen so that the probability of a true positive, P(R|H) in the two-class problem, is maximized while the probability of a false positive, P(R|F) in the two-class problem, is minimized (Ralston, 1998). However, there are a total of 2^N distinct possible Boolean fusion rules. When N is large, it is not feasible to test this many rules, but a smaller subset of all possible fusion

rules that represents the best performance can be defined and selected for a given sensor suite (Ralston, 1998).

When finding the subset of all possible Boolean fusion rules, there are two obvious rules: "never declare hostile" and "always declare hostile." The least conservative rule is "always declare hostile" where $R_j = 1$ for all j. The most conservative rule is "never declare hostile" where $R_j = 0$ for all j (Ralston, 1998). The next most conservative rule is the rule which includes just one state which has the highest likelihood ratio $P(S_j|H)/P(S_j|F)$. This fusion rule is better than any other fusion rule with just one single state included. The next fusion rule includes the previous rule as well as the next most likely state (i.e., the second rule includes two states). This fusion rule is better than any other fusion rule with just two states included. This process is repeated until the least conservative rule is reached or $R_j = 1$ for all j (Ralston, 1998). In essence, this method creates the ISOC boundary. The following ISOC boundary algorithm will create this boundary (Storm, Bauer, and Oxley, 2003).

1. Compute $P(S_j|T)$ for all j and T using data from the sensor probability matrices from the individual sensors.

2. Compute $LR^j = P(S_j H)/P(S_j|F)$ for all j, the likelihood ratio for all sensor output state combinations.

3. Rank LR^j for all j from highest to lowest, where $LR_{[1]}^{j_1}$ is the largest LR^j and $LR_{[N]}^{j_N}$ is the smallest LR^j, such that

$$LR_{[1]}^{j_1} > LR_{[2]}^{j_2} > \ldots > LR_{[N]}^{j_N}.$$

4. Choose S_j corresponding to the largest remaining $LR_{[N]}^{j_N}$ to be included in the fusion rule (i.e., $r_{j_N} = 1$ in R).

5. Go to 3 unless $r_j = 1$ for all j.

Using the data from the sensor probability matrices, the N distinct CIS configurations are tested and "turned on" in decreasing order of their likelihood ratios (Ralston, 1998). In a system with N states, there will be N+1 points that connect the most conservative rule and least conservative rule. Each of these points is a valid fusion rule; each rule provides an alternative trade-off between fratricide (incorrectly targeting a friendly) and effectiveness (correctly targeting a hostile). There is no rule that provides a higher level of effectiveness at the same or lower fratricide rate; there is no rule that can provide a lower level of fratricide at the same or higher level of effectiveness (Ralston, 1998). The optimal trade-off between fratricide and effectiveness depends on combat requirements (Ralston, 1998).

Optimal Rule Using Total Cost

Now that a subset of all possible rules has been identified, the optimal rule must be chosen. For each of the rules in the subset, a cost can be calculated. These costs depend only on the prior probabilities and the relative costs (Haspert, 2000). The cost equation is given by

$$C_{Total} = C_{False\ Negative} * P_{Hostile} * P(\text{False Negative}) + C_{False\ Positive} * P_{Friend} * P(\text{False Positive}).$$

where

C_{Total} = Total Cost

$C_{False\ Negative}$ = Cost of False Negative

$C_{False\ Positive}$ = Cost of False Positive

$P_{Hostile}$ = Prior Probability of a Hostile

P_{Friend} = Prior Probability of a Friend

P(True Positive) = P(R|H) = Probability of a True Positive

P(False Negative) = 1-P(R|H) = Probability of a False Negative

P(False Positive) = P(R|F) = Probability of a False Positive

P(False Negative) = 1 – P(True Positive) (Haspert, 2000).

Finally, the lowest cost rule can be chosen as the optimal rule. Figure 2 is a process diagram of the ISOC Fusion Method.

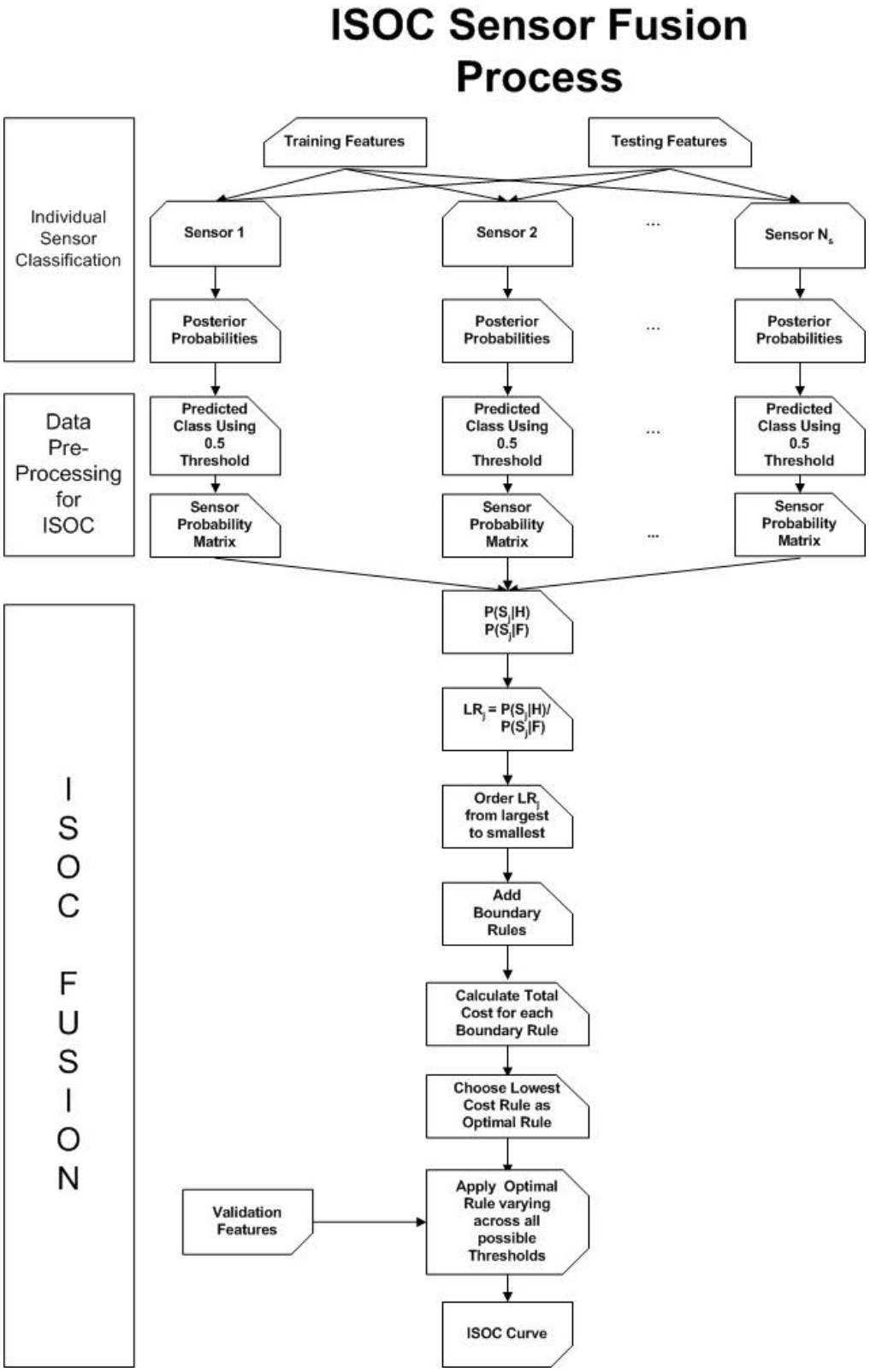


Figure 2: ISOC Sensor Fusion Process (Haspert, 2002).

ROC Fusion Methods

Two possible procedures for sensor fusion are called the ROC "Across" Fusion Method and the ROC "Within" Fusion Method. ROC "Across" Fusion is applicable when multiple sensors are monitoring multiple critical components in different feature sets. The ROC "Across" Method is concerned with the state of a collection of components as viewed by multiple sensors. On the other hand, ROC "Within" Fusion is applicable when multiple sensors are monitoring the same critical component in the same feature set. The ROC "Within" Method is concerned with the state of a single component as viewed by multiple sensors. In this thesis, only the ROC "Within" Method is used for sensor fusion.

ROC "Within" Fusion Method

While the ISOC method finds the optimal rule for a given threshold, the ROC "Within" Fusion method finds the optimal thresholds for each classifier for a given rule, the "logical or" rule. The "Within" Fusion method fuses the ROC curves together from individual sensors using the same or different feature sets to form a fused ROC curve (Clutz, 2000). Each individual classifier will output a sensor probability matrix shown in Table 1 where the definitions of true positive, false positive, true negative, and false negative are the same as above. Again, any indication "H" ("H"|H and "H"|F) is considered a positive and any indication "F" ("F"|H and "F"|F) is considered a negative. Let P_{TP}^{A} be the probability of true positive for classifier A, P_{FP}^{A} be the probability of false positive for classifier A, P_{TN}^{A} be the probability of true negative for classifier A, and P_{FN}^{A} be the probability of false negative for classifier A. Let P_{TP}^{B} be the probability of true

positive for classifier B, P_{FP}^{B} be the probability of false positive for classifier B, P_{TN}^{B} be the probability of true negative for classifier B, and P_{FN}^{B} be the probability of false negative for classifier B (Clutz, 2000). The ROC curve for each classifier is the set of coordinate points where a value of true positive (ordinate) is specified for every value of false positive (abscissa). Each of these coordinate points corresponds to a different threshold value for the individual classifier. The ROC "Within" Fusion Method uses these coordinate pairs, at common points along the abscissa for classifier A and classifier B, to form a new fused ROC curve (Clutz, 2000). Let classifier C be the classifier resulting from fusing classifiers A and B according to the "logical or" rule. Classifier C will result in a positive indication in three cases: when both classifier A and classifier B indicate positive, when only classifier A indicates positive, and when only classifier B indicates positive. For a two-class problem, $P_{FP} = 1 - P_{TN}$ which implies $P_{FP}^{C} = 1 - P_{TN}^{C}$. Assuming the logical "or" rule is used and assuming the independence of classifiers A and B, then

$$P_{FP}^{C} = 1 - (P_{TN}^{A} * P_{TN}^{B}) = 1 - (1 - P_{FP}^{A}) * (1 - P_{FP}^{B}) = (P_{FP}^{A} + P_{FP}^{B} - P_{FP}^{A} * P_{FP}^{B}).$$

For a two-class problem, $P_{TP} = 1 - P_{FN}$ so that $P_{TP}^{C} = 1 - P_{FN}^{C}$. Assuming independence of classifiers A and B, then (Clutz, 2000)

$$P_{TP}^{C} = 1 - P_{FN}^{A} * P_{FN}^{B} = 1 - (1 - P_{TP}^{A}) * (1 - P_{TP}^{B}) = (P_{TP}^{A} + P_{TP}^{B} - P_{TP}^{A} * P_{TP}^{B}).$$

Thus, the point on the fused ROC curve is given by the coordinate pair (Clutz, 2000)

$$(P_{FP}^{C}, P_{TP}^{C}) = (P_{FP}^{A} + P_{FP}^{B} - P_{FP}^{A} * P_{FP}^{B}, P_{TP}^{A} + P_{TP}^{B} - P_{TP}^{A} * P_{TP}^{B}).$$

Using these results, an optimization algorithm can be used to form the fused ROC curve and find the optimal thresholds for each individual classifier. Let p be a value of false positive for classifier A and $f_A(p)$ be a value of true positive for classifier A.

Similarly, let q be a value of false positive for classifier B and $f_B(q)$ be a value of true positive for classifier B. Let r^* be a value of the false positive for the fused classifier C and $f_C(r^*)$ be a value of true positive for the fused classifier C. It should be noted that q is a function of r and p; that is

$$r = p + q - p*q.$$

Thus, $q = Q(r, p) = \frac{(r - p)}{(1 - p)}$. Using this notation, the equation above can be rewritten as

$$(r\ , f_C(r\ )) = (p + q - p * q, \max_{0 \le p \le r}[f_A(p) + f_B(Q(r, p)) - f_A(p) * f_B(Q(r, p))]) \ .$$

Now, for each value of r, a value of p, denoted p^*, can be found such $f_A(p) + f_B(Q(r, p)) - f_A(p) * f_B(Q(r, p))$ is maximized subject to $0 \le p \le r$ (Storm, Bauer, and Oxley, 2003). After p^* is determined, $f_A(p^*)$ can be read from the ROC curve for classifier A. The optimal threshold for classifier A is the value θ^* that yields p* and $f_A(p^*)$. Using the relationship $r^* = p^* + q^* - p^* * q^*$, q^* can be determined, and $f_B(q^*)$ can be read from the ROC curve for classifier B. The optimal threshold for classifier B is the value φ^* that yields q^* and $f_B(q^*)$. This can be done for all values of r*. After thresholds for each classifier have been found for each value of r^*, these thresholds can be applied to an independent data set for validation. Figure 3 is a process diagram of the ROC "Within" OR Fusion Process.

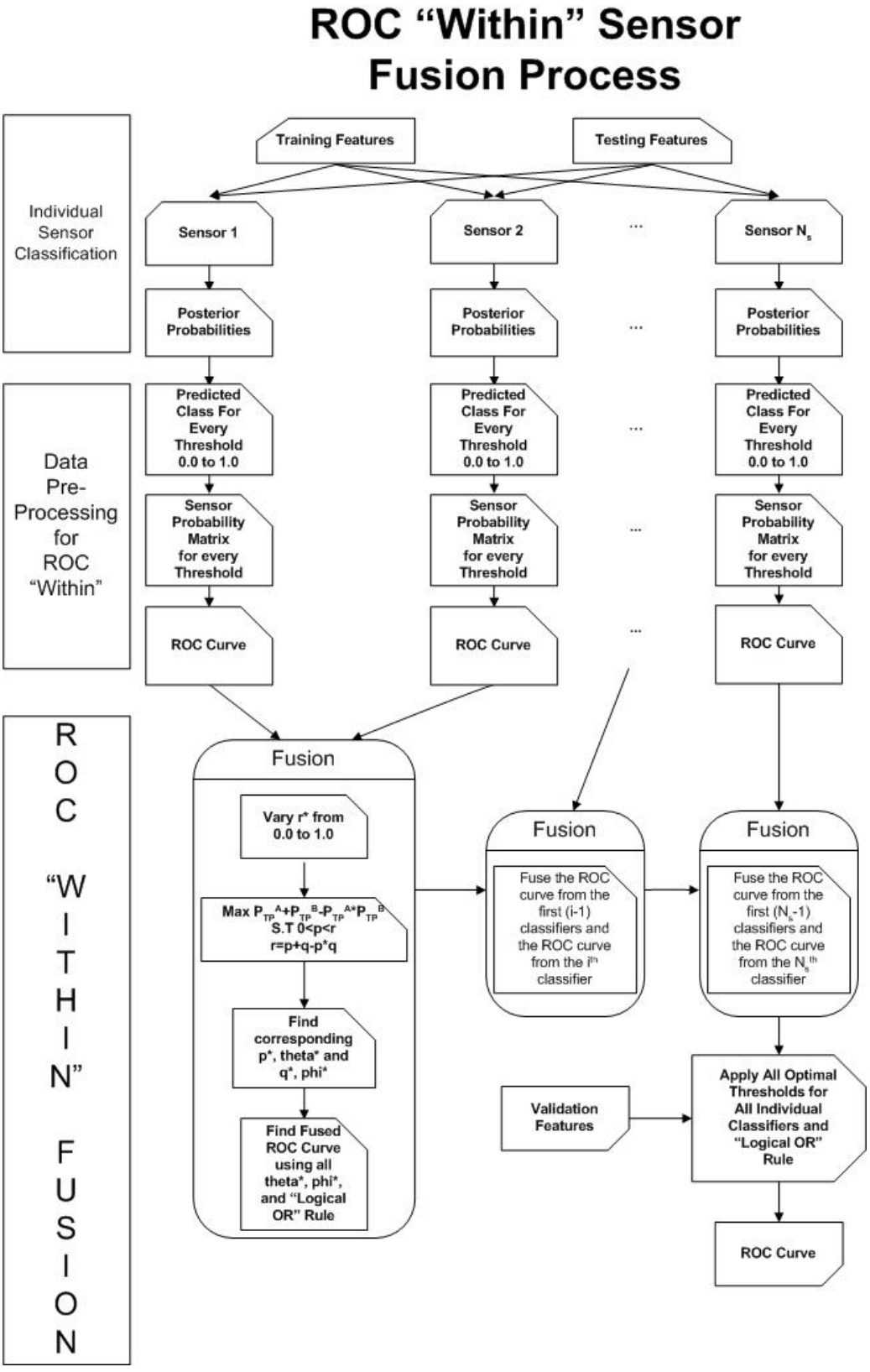


Figure 3: ROC "Within" OR Sensor Fusion Process.

PNN Fusion Method

The probabilistic neural network (PNN) fusion method is a simplistic fusion method that involves training a PNN on the posterior probabilities from the individual classifiers. The result is a single, fused classification. The PNN has been used successfully to solve a variety of classification problems (Wasserman and Nostrand, 1993). When compared to the standard back-propagation algorithm, the PNN has the following major advantages: rapid training, guaranteed convergence to a Bayesian Optimal Classifier with enough training data, allows deletion or addition from training data without retraining, and confidence indication on its output (Wasserman and Nostrand, 1993).

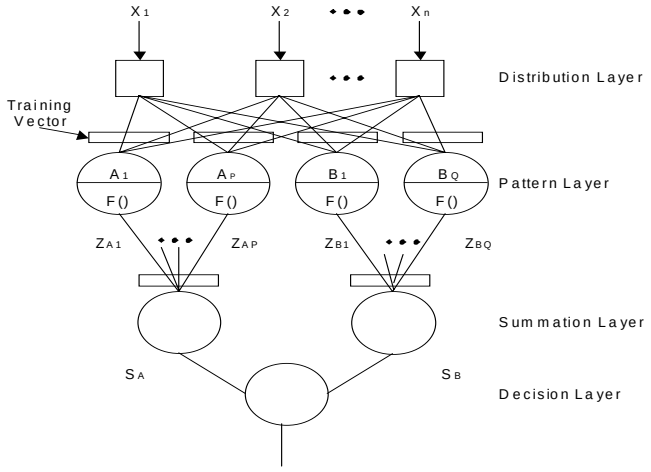


Figure 4: A Probabilistic Neural Network (Wasserman and Nostrand, 1993).

This method is based on the assumption that the feature sets are normalized and independent and identically distributed multivariate normal with common variance σ^2. The normalized input vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is applied to the distribution layer neurons. This input vector contains the features to be classified by the PNN. The distribution layer does not perform any calculations; it is simply a connection point

(Wasserman and Nostrand, 1993). Each training vector is used to calculate a set of weights where each weight has a value from a component of that vector. The pattern layer neurons are grouped together by the true classification of the associated training vector; these individual neurons sum the weighted inputs from the distribution layer neurons (Wasserman and Nostrand, 1993). This is equivalent to taking the sum of squares of the training set and the test set, $(X-X_{Ri})^T(X-X_{Ri})$ where X_{Ri} is the i^{th} exemplar in the R^{th} class from the training set. Because of the normalization, this reduces to $(X_{Ri}^T X_i - 1)$ (Wasserman and Nostrand, 1993). Then, the pattern layer neurons apply a nonlinear function to the corresponding sum. This produces the output $Z_{c,i}$, where c indicates the true class of the training vector and i indicates the pattern layer neuron (Wasserman and Nostrand, 1993). The nonlinear function for $Z_{c,i}$ is

$$Z_{c,i} = \exp[\frac{X_{Ri}^T X_i - 1}{\sigma^2}].$$

In this equation, **X** is defined above and the set of weights corresponding to a pattern neuron represent a training vector $\mathbf{X}_{Ri} = (X_{R1}, X_{R2}, \ldots, X_{Rn})$. The summation layer simply sums the Z_{ci} for each class (Wasserman and Nostrand, 1993). Thus, the output of the summation layer for a specific class, S_c is

$$S_c = \sum_{i=1} Z_{ci} .$$

The decision layer compares S_c for all classes and assigns the input vector to the class with the largest corresponding S_c. In essence, this PNN assigns a new feature set to the class that the feature set has the largest probability of being in under the multivariate normal distribution. A PNN can be extended to any number of classes by adding pattern layer neurons and a summation layer neuron for each class (Wasserman and Nostrand,

1993). Figure 5 is a process diagram of the PNN Fusion Process. The PNN Fusion process only uses half the data that the ISOC and ROC "Within" Fusion methods use.

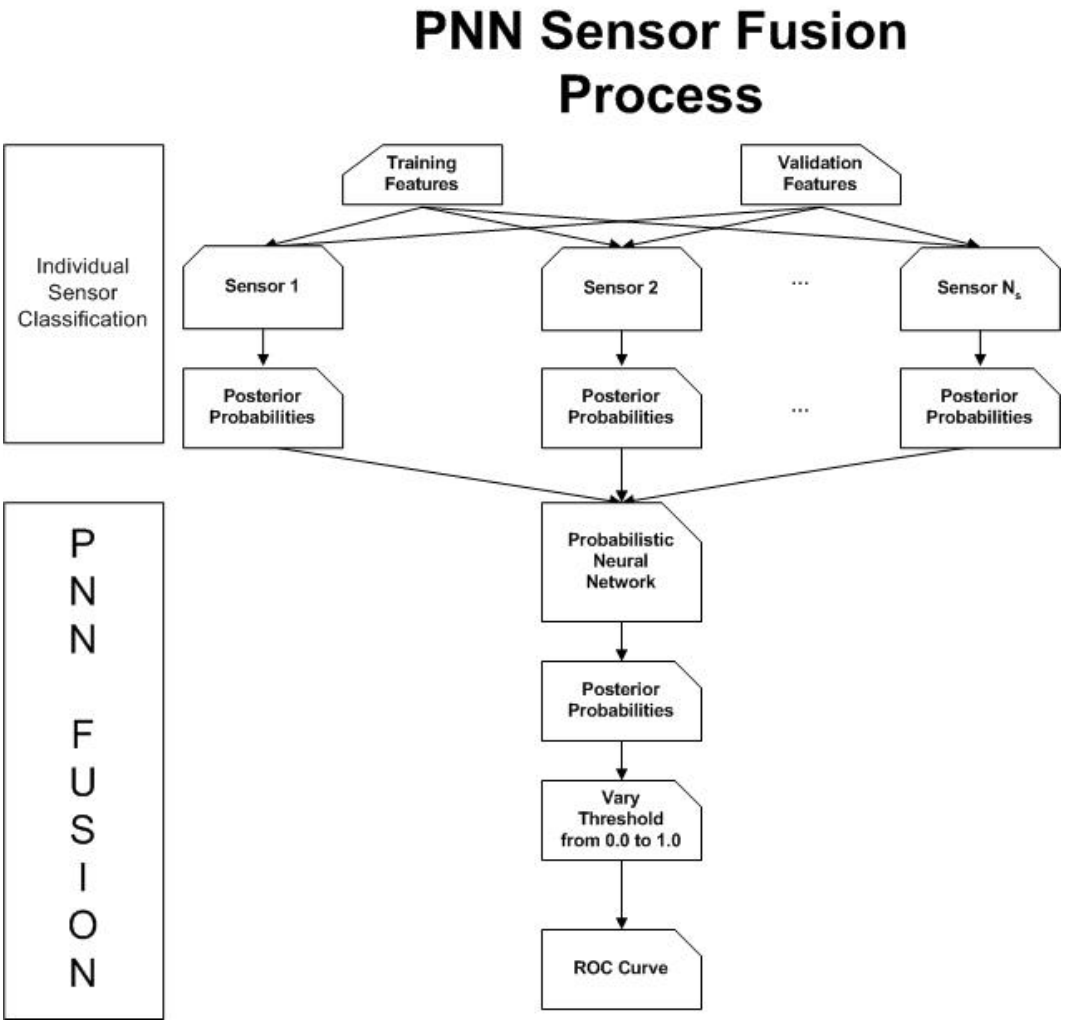


Figure 5: PNN Sensor Fusion Process.

Generalized Regression Neural Network (GRNN) Model

A GRNN has a very similar structure to a PNN, but it has one slight difference. While the PNN simply sums the nonlinear function for each class, the GRNN also sums this across all classes. Then, each S_c is divided by the sum of all the S_c values and that is the corresponding activation. Thus, all activations are standardized to a value between 0 and 1 (Wasserman and Nostrand, 1993).

Sample Size Considerations

In real world problems, distribution parameters are not known, and analysts are typically restrained to small training sets. The size of the training set, especially relative to the dimensionality of the problem or number of features used, will ultimately determine how close the estimated distribution parameters are to those of the true distribution. In other words, in a problem with few features, fewer training exemplars would be needed when compared to the requirements of a problem with many features. As the number of features grows, the sample size of the training set must also grow (Fukunaga and Hayes, 1989).

Sample size also plays a key part in comparing a linear classifier and quadratic classifier. If the covariance matrices for the two classes are equal and the true covariance matrix was used, the quadratic classifier and the linear classifier are the same. However, when the approximated covariance matrices are used, the approximations of the covariance matrices will not be the same even though their true covariance matrices yield the same results. Since the linear classifier will use all the data to calculate the covariance matrix, it will provide a better approximation of the true covariance matrix than that of the quadratic classifier. The quadratic classifier would need much more data to get as good of an approximation as the linear classifier. Thus, in a case where the covariance matrices are truly equal, the linear classifier is the more robust classifier (Fukunaga and Hayes, 1989).

Chapter Summary

This chapter summarized the important literature used to conduct this thesis. First, Air Force guidance on data fusion was summarized, and the statistical

independence assumption was explored. Next, the four fusion methods employed in this thesis were described in detail. Finally, some sample size considerations were discussed.

III. Methodology

Introduction

This chapter lays out the basic methodology used in this thesis. First, it describes the different types of correlation introduced into the fusion models. Next, it describes the data generation process for each of the different problems explored. Next, the general experimental design is discussed and some application issues for each of the four fusion methods are detailed. Finally, feature selection, Total Probability of Misclassification (TPM), and sample size variation are discussed.

Correlation

In this thesis, multiple feature sets, each containing multiple features, are generated for experimentation. Some level of correlation is present among these features. Two types of correlation are considered: inter-correlation and intra-correlation. The first type of correlation considered is inter-correlation; this is the correlation between features in a given data set. Figure 6 is a notional diagram representing inter-correlation.

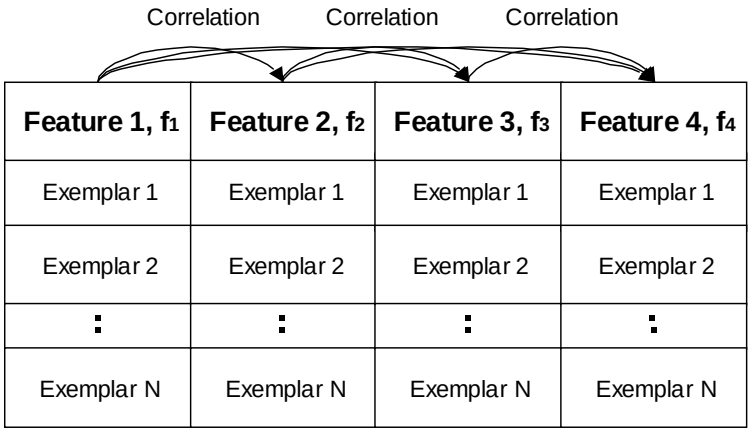


Figure 6: Inter-correlation.

The second type of correlation considered is intra-correlation or autocorrelation; this is the correlation between observations in a single feature. Figure 7 is a notional diagram representing intra-correlation.

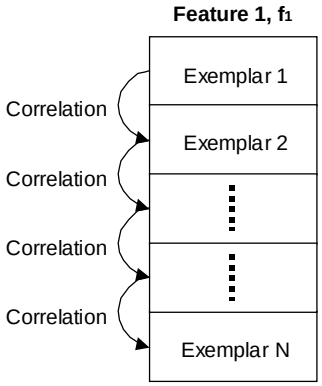
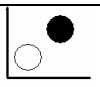
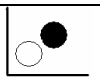
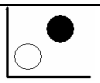
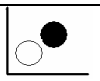
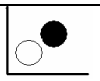
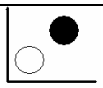


Figure 7: Intra-correlation.

Data Generation

Since real-world data is not available, data was generated for a variety of problems for analysis for this thesis. The following Table 4 summarizes the problems analyzed in this thesis.

Table 4: Data Generation Descriptions.

Problem #	Problem Name	Problem Description
1 	4 Feature Case	Recreates Storm work; average ROC curve of 5 runs as response
2 	8 Feature Case	Adds noise and redundant features to problem 1; changes mean of class 1
3 	8 Feature with Autocorrelation Case	Adds autocorrelation to problem 2; changes mean of class 1
4 	8 Feature Triangle Case	Changes geometry of problem 2
5 	8 Feature XOR Case	Changes geometry of problem 4
6 	8 Feature XOR with Autocorrelation Case	Adds autocorrelation to problem 5
7 	20 Feature with Feature Selection Case	Adds more noise and redundant features to problem 2; explores only 2 sample sizes; changes mean of class 1
8 	36 Feature with Feature Selection Case	Adds more noise and redundant features to problem 7; explores only 1 sample size and 1 correlation level
9 	TPM Exploration	Examines problem 1 at 3 specific levels of correlation

Problem 1: 4 Feature Case 

Let $F = F_1 \times F_2 \subset R^4$ where $F_1 \subset R^2$ is the feature set observed by sensor 1, a linear discriminant function, and $F_2 \subset R^2$ is the feature set observed by sensor 2, a quadratic discriminant function. The correlation of the data is given by

$$\Sigma^i = \begin{bmatrix} \Sigma^i_{F_1,F_1} & \Sigma^i_{F_1,F_2} \\ \Sigma^i_{F_2,F_1} & \Sigma^i_{F_2,F_2} \end{bmatrix}.$$

Since all the features in the individual feature sets are statistically independent,

$\Sigma^i_{F_1,F_1} = \Sigma^i_{F_2,F_2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Also, $\Sigma^i_{F_1,F_2} = \Sigma^i_{F_2,F_1} = \begin{bmatrix} 0 & \rho \\ \rho & 0 \end{bmatrix}$ where $\rho \in \{0.0,0.2,0.4,0.6,0.8,0.9\}$ and $\Sigma^i_{F_j,F_k}$ is the correlation matrix between the features contained in feature set j and feature set k in class i (i = 0,1; j, k = 1,2). Now, let F_1^0 be the features from feature set 1 in class 0 and F_1^1 be the features from feature set 1 in class 1 where $F_1 = F_1^0 \cup F_1^1$. Let μ_1^0 be the mean of feature set 1 in class 0 and μ_1^1 be the mean of feature set 1 in class 1. Let $F_1^0 \sim N(\mu_1^0, \Sigma^0_{F_1,F_1})$ and $F_1^1 \sim N(\mu_1^1, \Sigma^1_{F_1,F_1})$ where $\mu_1^0 = (0,0)^T$ and $\mu_1^1 = (0.95,0.95)^T$. Let $F_2 = F_2^0 \cup F_2^1$ where $F_2^0 \sim N(\mu_2^0, \Sigma^0_{F_2,F_2})$ and $F_2^1 \sim N(\mu_2^1, \Sigma^1_{F_2,F_2})$ where $\mu_2^0 = (0,0)^T$ and $\mu_2^1 = (1.15,1.15)^T$.

Problem 2: 8 Feature Case 

Let $F = F_1 \times F_2 \subset R^8$ where $F_1 \subset R^4$ is the feature set observed by sensor 1, a linear discriminant function, and $F_2 \subset R^4$ is the feature set observed by sensor 2, a quadratic discriminant function. The correlation of the data is given by

$$\Sigma^i = \begin{bmatrix} \Sigma^i_{F_1,F_1} & \Sigma^i_{F_1,F_2} \\ \Sigma^i_{F_2,F_1} & \Sigma^i_{F_2,F_2} \end{bmatrix}.$$

In this case, each feature set will contain 2 independent features (separated in mean), 1 redundant feature, and 1 noise feature (same mean).

$$\Sigma^i_{F_1,F_1} = \Sigma^i_{F_2,F_2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \rho_{red} & 0 \\ 0 & \rho_{red} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Also,

$$\Sigma^{i}_{F_1,F_2} = \Sigma^{i}_{F_2,F_1} = \begin{bmatrix} 0 & \rho & \rho_{ind} & 0 \\ \rho & 0 & 0 & 0 \\ \rho_{ind} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where $\rho \in \{0.0,0.2,0.4,0.6,0.8,0.9\}$, $\rho_{red} = 0.95$ is the correlation level of the redundant feature, and $\rho_{ind} = \rho * \rho_{red}$ is the correlation level induced by ρ and ρ_{red}. $\Sigma^{i}_{F_j,F_k}$ is the correlation matrix between the features contained in feature set j and feature set k in class i (i = 0,1; j, k = 1,2). Now, let F_1^0 be the features from feature set 1 in class 0 and F_1^1 be the features from feature set 1 in class 1 where $F_1 = F_1^0 \cup F_1^1$. Let μ_1^0 be the mean of feature set 1 in class 0 and μ_1^1 be the mean of feature set 1 in class 1. Let $F_1^0 \sim N(\mu_1^0, \Sigma^0_{F_1,F_1})$ and $F_1^1 \sim N(\mu_1^1, \Sigma^1_{F_1,F_1})$ where $\mu_1^0 = (0,0,0,0)^T$ and $\mu_1^1 = (0.50,0.50,0.50,0)^T$. Let $F_2 = F_2^0 \cup F_2^1$ where $F_2^0 \sim N(\mu_2^0, \Sigma^0_{F_2,F_2})$ and $F_2^1 \sim N(\mu_2^1, \Sigma^1_{F_2,F_2})$ where $\mu_2^0 = (0,0,0,0)^T$ and $\mu_2^1 = (0.75,0.75,0.75,0)^T$.

Problem 3: 8 Feature with Autocorrelation Case 

Let $F = F_1 \times F_2 \subset R^8$ where $F_1 \subset R^4$ is the feature set observed by sensor 1, a linear discriminant function, and $F_2 \subset R^4$ is the feature set observed by sensor 2, a quadratic discriminant function. The correlation of the data is given by

$$\Sigma^{i} = \begin{bmatrix} \Sigma^{i}_{F_1,F_1} & \Sigma^{i}_{F_1,F_2} \\ \Sigma^{i}_{F_2,F_1} & \Sigma^{i}_{F_2,F_2} \end{bmatrix}.$$

In this case, each feature set will contain 2 independent features (separated in mean), 1 redundant feature, and 1 noise feature (same mean).

$$\Sigma^{i}_{F_1,F_1} = \Sigma^{i}_{F_2,F_2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \rho_{red} & 0 \\ 0 & \rho_{red} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Also,

$$\Sigma^{i}_{F_1,F_2} = \Sigma^{i}_{F_2,F_1} = \begin{bmatrix} 0 & \rho & \rho_{ind} & 0 \\ \rho & 0 & 0 & 0 \\ \rho_{ind} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where $\rho \in \{0.0,0.2,0.4,0.6,0.8,0.9\}$, $\rho_{red} = 0.95$ is the correlation level of the redundant feature, and $\rho_{ind} = \rho * \rho_{red}$ is the correlation level induced by ρ and ρ_{red}. $\Sigma^{i}_{F_j,F_k}$ is the correlation matrix between the features contained in feature set j and feature set k in class i (i = 0,1; j, k = 1,2). Now, let F_1^0 be the features from feature set 1 in class 0 and F_1^1 be the features from feature set 1 in class 1 where $F_1 = F_1^0 \cup F_1^1$. Let μ_1^0 be the mean of feature set 1 in class 0 and μ_1^1 be the mean of feature set 1 in class 1. Let $F_1^0 \sim N(\mu_1^0, \Sigma^0_{F_1,F_1})$ and $F_1^1 \sim N(\mu_1^1, \Sigma^1_{F_1,F_1})$ where $\mu_1^0 = (0,0,0,0)^T$ and $\mu_1^1 = (0.95,0.95,0.95,0)^T$. Let $F_2 = F_2^0 \cup F_2^1$ where $F_2^0 \sim N(\mu_2^0, \Sigma^0_{F_2,F_2})$ and $F_2^1 \sim N(\mu_2^1, \Sigma^1_{F_2,F_2})$ where $\mu_2^0 = (0,0,0,0)^T$ and $\mu_2^1 = (1.15,1.15,1.15,0)^T$. This adds the appropriate level of correlation between feature sets. In addition, $\rho_{auto} \in \{0.0,0.5,0.9\}$ is the level of autocorrelation within a feature set. Let $z(t) \subset R^8, t \in \{1,2,...,N\}$ where N is the number of training exemplars, be one exemplar in the feature space where $z(t) \sim N(\bar{0}, \Sigma^0)$; it is one row of the matrix of features described above. Let $A = \rho_{auto} * I$, $B = (\sqrt{1-\rho^2_{auto}}) * I$, and $\varepsilon(t) \sim N(\bar{0}, (B * \Sigma^0 * B))$ for each t. Then,

$z(t) = A * z(t-1) + \varepsilon(t)$ (Laine, 2003). Once the appropriate number of exemplars has been generated, the means can be added to the corresponding classes.

Problem 4: 8 Feature Triangle Case 

Every problem up to this point is a fairly simple problem separating two multivariate normal populations. The Triangle problem is a slightly more complicated problem building toward the XOR problem. It is interesting to see how each of the fusion methods will perform in the face of this more complicated problem. Each class will contain two multivariate populations; thus, four multivariate populations will be generated. Two will be assigned to one class and two to the other class. All four multivariate distributions will have the same covariance structure. Let $F = F_1 \times F_2 \subset R^8$ where $F_1 \subset R^4$ is the feature set observed by sensor 1, a linear discriminant function, and $F_2 \subset R^4$ is the feature set observed by sensor 2, a quadratic discriminant function. The correlation of the data is given by

$$\Sigma^i = \begin{bmatrix} \Sigma^i_{F_1,F_1} & \Sigma^i_{F_1,F_2} \\ \Sigma^i_{F_2,F_1} & \Sigma^i_{F_2,F_2} \end{bmatrix}.$$

In this case, each feature set will contain 2 independent features (separated in mean), 1 redundant feature, and 1 noise feature (same mean).

$$\Sigma^i_{F_1,F_1} = \Sigma^i_{F_2,F_2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \rho_{red} & 0 \\ 0 & \rho_{red} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Also,

$$\Sigma^{i}_{F_1,F_2} = \Sigma^{i}_{F_2,F_1} = \begin{bmatrix} 0 & \rho & \rho_{ind} & 0 \\ \rho & 0 & 0 & 0 \\ \rho_{ind} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where $\rho \in \{0.0,0.2,0.4,0.6,0.8,0.9\}$, $\rho_{red} = 0.95$ is the correlation level of the redundant feature, and $\rho_{ind} = \rho * \rho_{red}$ is the correlation level induced by ρ and ρ_{red}. $\Sigma^{i}_{F_j,F_k}$ is the correlation matrix between the features contained in feature set j and feature set k in class i (i = 0,1; j, k = 1,2). Now, let F_1^{01} be the first set of features from feature set 1 in class 0, F_1^{02} be the second set of features from feature set 1 in class 0, F_1^{11} be the first set of features from feature set 1 in class 1, and F_1^{12} be the second set of features from feature set 1 in class 1 where $F_1^0 = F_1^{01} \cup F_1^{02}$, $F_1^1 = F_1^{11} \cup F_1^{12}$ and $F_1 = F_1^0 \cup F_1^1$. Let μ_1^{01} be the mean of the first set of features in feature set 1 in class 0, μ_1^{02} be the mean of the second set of features in feature set 1 in class 0, μ_1^{11} be the mean of the first set of features in feature set 1 in class 1, and μ_1^{12} be the mean of the second set of features in feature set 1 in class 1. Let $F_1^{01} \sim N(\mu_1^{01}, \Sigma^0_{F_1,F_1})$, $F_1^{02} \sim N(\mu_1^{02}, \Sigma^0_{F_1,F_1})$, $F_1^{11} \sim N(\mu_1^{11}, \Sigma^1_{F_1,F_1})$, and $F_1^{12} \sim N(\mu_1^{12}, \Sigma^1_{F_1,F_1})$ where $\mu_1^{01} = (0,0,0,0)^T$, $\mu_1^{02} = (0.95,0.95,0.95,0)^T$, $\mu_1^{11} = (0.95,0,0,0)^T$ and $\mu_1^{12} = (0.95,0,0,0)^T$.

Let $F_2^0 = F_2^{01} \cup F_2^{02}$, $F_2^1 = F_2^{11} \cup F_2^{12}$, and $F_2 = F_2^0 \cup F_2^1$ where $F_2^{01} \sim N(\mu_2^{01}, \Sigma^0_{F_1,F_1})$, $F_2^{02} \sim N(\mu_2^{02}, \Sigma^0_{F_1,F_1})$, $F_2^{11} \sim N(\mu_2^{11}, \Sigma^1_{F_1,F_1})$, and $F_2^{12} \sim N(\mu_2^{12}, \Sigma^1_{F_1,F_1})$ where $\mu_2^{01} = (0,0,0,0)^T$, $\mu_2^{02} = (1.15,1.15,1.15,0)^T$, $\mu_2^{11} = (1.15,0,0,0)^T$, and $\mu_2^{12} = (1.15,0,0,0)^T$.

Problem 5: 8 Feature XOR Case 

Every problem up to this point is a fairly simple problem separating two multivariate normal populations. The XOR problem is a more complicated problem. It is interesting to see how each of the fusion methods will perform in the face of this more complicated problem. Each class will contain two multivariate populations; thus, four multivariate populations will be generated. Two will be assigned to one class and two to the other class. All four multivariate distributions will have the same covariance structure. Let $F = F_1 \times F_2 \subset R^8$ where $F_1 \subset R^4$ is the feature set observed by sensor 1, a linear discriminant function, and $F_2 \subset R^4$ is the feature set observed by sensor 2, a quadratic discriminant function. The correlation of the data is given by

$$\Sigma^i = \begin{bmatrix} \Sigma^i_{F_1,F_1} & \Sigma^i_{F_1,F_2} \\ \Sigma^i_{F_2,F_1} & \Sigma^i_{F_2,F_2} \end{bmatrix}.$$

In this case, each feature set will contain 2 independent features (separated in mean), 1 redundant feature, and 1 noise feature (same mean).

$$\Sigma^i_{F_1,F_1} = \Sigma^i_{F_2,F_2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \rho_{red} & 0 \\ 0 & \rho_{red} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Also,

$$\Sigma^i_{F_1,F_2} = \Sigma^i_{F_2,F_1} = \begin{bmatrix} 0 & \rho & \rho_{ind} & 0 \\ \rho & 0 & 0 & 0 \\ \rho_{ind} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where $\rho \in \{0.0,0.2,0.4,0.6,0.8,0.9\}$, $\rho_{red} = 0.95$ is the correlation level of the redundant feature, and $\rho_{ind} = \rho * \rho_{red}$ is the correlation level induced by ρ and ρ_{red}. $\Sigma^{i}_{F_j,F_k}$ is the correlation matrix between the features contained in feature set j and feature set k in class i (i = 0,1; j, k = 1,2). Now, let F_1^{01} be the first set of features from feature set 1 in class 0, F_1^{02} be the second set of features from feature set 1 in class 0, F_1^{11} be the first set of features from feature set 1 in class 1, and F_1^{12} be the second set of features from feature set 1 in class 1 where $F_1^0 = F_1^{01} \cup F_1^{02}$, $F_1^1 = F_1^{11} \cup F_1^{12}$ and $F_1 = F_1^0 \cup F_1^1$. Let μ_1^{01} be the mean of the first set of features in feature set 1 in class 0, μ_1^{02} be the mean of the second set of features in feature set 1 in class 0, μ_1^{11} be the mean of the first set of features in feature set 1 in class 1, and μ_1^{12} be the mean of the second set of features in feature set 1 in class 1. Let $F_1^{01} \sim N(\mu_1^{01}, \Sigma^0_{F_1,F_1})$, $F_1^{02} \sim N(\mu_1^{02}, \Sigma^0_{F_1,F_1})$, $F_1^{11} \sim N(\mu_1^{11}, \Sigma^1_{F_1,F_1})$, and $F_1^{12} \sim N(\mu_1^{12}, \Sigma^1_{F_1,F_1})$ where $\mu_1^{01} = (0,0,0,0)^T$, $\mu_1^{02} = (0.95,0.95,0.95,0)^T$, $\mu_1^{11} = (0,0.95,0.95,0)^T$ and $\mu_1^{12} = (0.95,0,0,0)^T$.

Let $F_2^0 = F_2^{01} \cup F_2^{02}$, $F_2^1 = F_2^{11} \cup F_2^{12}$, and $F_2 = F_2^0 \cup F_2^1$ where $F_2^{01} \sim N(\mu_2^{01}, \Sigma^0_{F_1,F_1})$, $F_2^{02} \sim N(\mu_2^{02}, \Sigma^0_{F_1,F_1})$, $F_2^{11} \sim N(\mu_2^{11}, \Sigma^1_{F_1,F_1})$, and $F_2^{12} \sim N(\mu_2^{12}, \Sigma^1_{F_1,F_1})$ where $\mu_2^{01} = (0,0,0,0)^T$, $\mu_2^{02} = (1.15,1.15,1.15,0)^T$, $\mu_2^{11} = (0,1.15,1.15,0)^T$, and $\mu_2^{12} = (1.15,0,0,0)^T$.

Problem 6: 8 Feature XOR with Autocorrelation Case 

This problem adds autocorrelation to the 8 feature XOR Problem. Again, each class will contain two multivariate populations; thus, four multivariate populations will

be generated. Two will be assigned to one class and two to the other class. All four multivariate distributions will have the same covariance structure. Let $F = F_1 \times F_2 \subset R^8$ where $F_1 \subset R^4$ is the feature set observed by sensor 1, a linear discriminant function, and $F_2 \subset R^4$ is the feature set observed by sensor 2, a quadratic discriminant function. The correlation of the data is given by

$$\Sigma^i = \begin{bmatrix} \Sigma^i_{F_1,F_1} & \Sigma^i_{F_1,F_2} \\ \Sigma^i_{F_2,F_1} & \Sigma^i_{F_2,F_2} \end{bmatrix}.$$

In this case, each feature set will contain 2 independent features (separated in mean), 1 redundant feature, and 1 noise feature (same mean).

$$\Sigma^i_{F_1,F_1} = \Sigma^i_{F_2,F_2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \rho_{red} & 0 \\ 0 & \rho_{red} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Also,

$$\Sigma^i_{F_1,F_2} = \Sigma^i_{F_2,F_1} = \begin{bmatrix} 0 & \rho & \rho_{ind} & 0 \\ \rho & 0 & 0 & 0 \\ \rho_{ind} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where $\rho \in \{0.0,0.2,0.4,0.6,0.8,0.9\}$, $\rho_{red} = 0.95$ is the correlation level of the redundant feature, and $\rho_{ind} = \rho * \rho_{red}$ is the correlation level induced by ρ and ρ_{red}. $\Sigma^i_{F_j,F_k}$ is the correlation matrix between the features contained in feature set j and feature set k in class i (i = 0,1; j, k = 1,2). Now, let F_1^{01} be the first set of features from feature set 1 in class 0, F_1^{02} be the second set of features from feature set 1 in class 0, F_1^{11} be the first set of features from feature set 1 in class 1, and F_1^{12} be the second set of features from feature

set 1 in class 1 where $F_1^0 = F_1^{01} \cup F_1^{02}, F_1^1 = F_1^{11} \cup F_1^{12}$ and $F_1 = F_1^0 \cup F_1^1$. Let μ_1^{01} be the mean of the first set of features in feature set 1 in class 0, μ_1^{02} be the mean of the second set of features in feature set 1 in class 0, μ_1^{11} be the mean of the first set of features in feature set 1 in class 1, and μ_1^{12} be the mean of the second set of features in feature set 1 in class 1. Let $F_1^{01} \sim N(\mu_1^{01}, \Sigma^0_{F_1,F_1})$, $F_1^{02} \sim N(\mu_1^{02}, \Sigma^0_{F_1,F_1})$, $F_1^{11} \sim N(\mu_1^{11}, \Sigma^1_{F_1,F_1})$, and $F_1^{12} \sim N(\mu_1^{12}, \Sigma^1_{F_1,F_1})$ where $\mu_1^{01} = (0,0,0,0)^T$, $\mu_1^{02} = (0.95,0.95,0.95,0)^T$, $\mu_1^{11} = (0,0.95,0.95,0)^T$ and $\mu_1^{12} = (0.95,0,0,0)^T$.

Let $F_2^0 = F_2^{01} \cup F_2^{02}$, $F_2^1 = F_2^{11} \cup F_2^{12}$, and $F_2 = F_2^0 \cup F_2^1$ where $F_2^{01} \sim N(\mu_2^{01}, \Sigma^0_{F_1,F_1})$, $F_2^{02} \sim N(\mu_2^{02}, \Sigma^0_{F_1,F_1})$, $F_2^{11} \sim N(\mu_2^{11}, \Sigma^1_{F_1,F_1})$, and $F_2^{12} \sim N(\mu_2^{12}, \Sigma^1_{F_1,F_1})$ where $\mu_2^{01} = (0,0,0,0)^T$, $\mu_2^{02} = (1.15,1.15,1.15,0)^T$, $\mu_2^{11} = (0,1.15,1.15,0)^T$, and $\mu_2^{12} = (1.15,0,0,0)^T$. This adds the appropriate level of correlation between feature sets. In addition, $\rho_{auto} \in \{0.0,0.5,0.9\}$ is the level of autocorrelation within a feature set. Let $z_{01}(t) \subset R^8, t \in \{1,2,...,N\}$ where N is the number of training exemplars, be one exemplar in the feature space for the first set of features in class 0 where $z_{01}(t) \sim N(\bar{0}, \Sigma^0)$ for all t; it is one row of the matrix of features described above. Let $z_{02}(t) \subset R^8, t \in \{1,2,...,N\}$ where N is the number of training exemplars, be one exemplar in the feature space for the second set of features in class 0 where $z_{02}(t) \sim N(\bar{0}, \Sigma^0)$ for all t; it is one row of the matrix of features described above. Let $z_{11}(t) \subset R^8, t \in \{1,2,...,N\}$ where N is the number of training exemplars, be one exemplar in the feature space for the first set of features in class 1 where

$z_{11}(t) \sim N(\bar{0}, \Sigma^0)$ for all t; it is one row of the matrix of features described above. Let $z_{12}(t) \subset R^8, t \in \{1,2,...,N\}$ where *N* is the number of training exemplars, be one exemplar in the feature space for the second set of features in class 1 where $z_{12}(t) \sim N(\bar{0}, \Sigma^0)$ for all t; it is one row of the matrix of features described above. Let $A = \rho_{auto} * I$, $B = (\sqrt{1-\rho_{auto}^2}) * I$, $\varepsilon_{01}(t) \sim N(\bar{0}, (B * \Sigma^0 * B))$, $\varepsilon_{02}(t) \sim N(\bar{0}, (B * \Sigma^0 * B))$, $\varepsilon_{11}(t) \sim N(\bar{0}, (B * \Sigma^0 * B))$, and $\varepsilon_{12}(t) \sim N(\bar{0}, (B * \Sigma^0 * B))$. Then, $z_{01}(t) = A * z_{01}(t-1) + \varepsilon_{01}(t)$, $z_{02}(t) = A * z_{02}(t-1) + \varepsilon_{02}(t)$, $z_{11}(t) = A * z_{11}(t-1) + \varepsilon_{11}(t)$, and $z_{12}(t) = A * z_{12}(t-1) + \varepsilon_{12}(t)$ (Laine, 2003). Once the appropriate number of exemplars has been generated, the means can be added to the corresponding populations, and the populations can be grouped together into the appropriate classes. Class 0 is composed of $z_{01}(t)$ and $z_{02}(t)$; Class 1 is composed of $z_{11}(t)$ and $z_{12}(t)$.

Problem 7: 20 Feature with Feature Selection Case 

Let $F = F_1 \times F_2 \subset R^{20}$ where $F_1 \subset R^{10}$ is the feature set observed by sensor 1 and $F_2 \subset R^{10}$ is the feature set observed by sensor 2. The correlation of the data is given by

$$\Sigma^i = \begin{bmatrix} \Sigma^i_{F_1,F_1} & \Sigma^i_{F_1,F_2} \\ \Sigma^i_{F_2,F_1} & \Sigma^i_{F_2,F_2} \end{bmatrix}.$$

In this case, each feature set will contain 2 independent features (separated in mean), 4 redundant features, and 4 noise features (same mean).

$$\Sigma^{i}_{F_1,F_1} = \Sigma^{i}_{F_2,F_2} = \left[\begin{array}{cccccc|cccc} 1 & 0 & \rho_{red1} & \rho_{red2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \rho_{red3} & \rho_{red4} & 0 & 0 & 0 & 0 \\ \rho_{red1} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \rho_{red2} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho_{red3} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho_{red4} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

Also,

$$\Sigma^{i}_{F_1,F_2} = \Sigma^{i}_{F_2,F_1} = \left[\begin{array}{cccccc|cccc} 0 & \rho & 0 & 0 & \rho_{ind3} & \rho_{ind4} & 0 & 0 & 0 & 0 \\ \rho & 0 & \rho_{ind1} & \rho_{ind2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho_{ind1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho_{ind2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \rho_{ind3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \rho_{ind4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

where $\rho \in \{0.0,0.2,0.4,0.6,0.8,0.9\}$, $\rho_{red1} = 0.60, \rho_{red2} = 0.50, \rho_{red3} = 0.40, \rho_{red4} = 0.30,$ are the correlation levels of the redundant features, and $\rho_{ind_m} = \rho * \rho_{red_m}$ are the correlation levels induced by ρ and ρ_{red_m} for all m=1,2,3,4. $\Sigma^{i}_{F_j,F_k}$ is the correlation matrix between the features contained in feature set j and feature set k in class i (i = 0,1; j, k = 1,2). Now, let F_1^0 be the features from feature set 1 in class 0 and F_1^1 be the features from feature set 1 in class 1 where $F_1 = F_1^0 \cup F_1^1$. Let μ_1^0 be the mean of feature set 1 in class 0 and μ_1^1 be the mean of feature set 1 in class 1. Let $F_1^0 \sim N(\mu_1^0, \Sigma^0_{F_1,F_1})$ and

$F_1^1 \sim N(\mu_1^1, \Sigma_{F_1,F_1}^1)$ where $\mu_1^0 = (0,0,0,0,0,0,0,0,0,0)^T$ and

$\mu_1^1 = (0.5,0.5,0.5,0.5,0.5,0.5,0,0,0,0)^T$. Let $F_2 = F_2^0 \cup F_2^1$ where $F_2^0 \sim N(\mu_2^0, \Sigma_{F_2,F_2}^0)$ and

$F_2^1 \sim N(\mu_2^1, \Sigma_{F_2,F_2}^1)$ where $\mu_2^0 = (0,0,0,0,0,0,0,0,0,0)^T$ and

$\mu_2^1 = (0.75,0.75,0.75,0.75,0.75,0.75,0,0,0,0)^T$. Two different sample sizes were used for this problem: 50 exemplars in each class and 1000 exemplars in each class. Three data sets were generated for each sample size and used as shown in the process flow diagrams above.

In this problem, after the data was generated as described above, feature selection was performed. To perform feature selection, discriminant loadings were calculated for each feature. In this project, the loading is defined to be the correlation between the feature and the posterior probability of being in class 1. After the loadings were calculated, any feature with a loading greater than 0.45 was kept as a good feature. In the small sample size problem, there were cases where none of the loadings were larger than 0.45. In these cases, only the feature with the largest loading was considered a good feature and kept for the remainder of the analysis. Then, discriminant analysis and classifier fusion were redone using only those good features. For comparison, analysis was also done without feature selection where all the features were kept and used for the discriminant analysis and sensor fusion. These two results were compared. This process was completed for each of the six correlation levels, each of the two sample sizes (50 exemplars in each class and 1000 exemplars in each class), with and without feature selection, over 15 runs; there were a total of 360 runs in this first experiment.

Problem 8: 36 Feature Case with Feature Selection Case 

Let $F = F_1 \times F_2 \subset R^{36}$ where $F_1 \subset R^{18}$ is the feature set observed by sensor 1 and $F_2 \subset R^{18}$ is the feature set observed by sensor 2. The correlation of the data is given by

$$\Sigma^i = \begin{bmatrix} \Sigma^i_{F_1,F_1} & \Sigma^i_{F_1,F_2} \\ \Sigma^i_{F_2,F_1} & \Sigma^i_{F_2,F_2} \end{bmatrix}.$$

In this case, each feature set will contain 2 independent features (separated in mean), 8 redundant features, and 8 noise features (same mean).

$$\Sigma^i_{F_1,F_1} = \Sigma^i_{F_2,F_2} = \begin{bmatrix} M_{11} & 0 \\ 0 & I \end{bmatrix}$$

where

$$M_{11} = \begin{bmatrix} 1 & 0 & \rho_{red1} & \rho_{red2} & \rho_{red3} & \rho_{red4} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \rho_{red5} & \rho_{red6} & \rho_{red7} & \rho_{red8} \\ \rho_{red1} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \rho_{red2} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \rho_{red3} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \rho_{red4} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & \rho_{red5} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & \rho_{red6} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & \rho_{red7} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & \rho_{red8} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

.

Also, $\Sigma^i_{F_1,F_2} = \Sigma^i_{F_2,F_1} = \begin{bmatrix} M_{12} & 0 \\ 0 & 0 \end{bmatrix}$

$$M_{12} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \rho_{ind5} & \rho_{ind6} & \rho_{ind7} & \rho_{ind8} \\ 0 & 1 & \rho_{ind1} & \rho_{ind2} & \rho_{ind3} & \rho_{ind4} & 0 & 0 & 0 & 0 \\ 0 & \rho_{ind1} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho_{ind2} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho_{ind3} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho_{ind4} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \rho_{ind5} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \rho_{ind6} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \rho_{ind7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \rho_{ind8} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where $\rho = 0.8$,

$\rho_{red1} = \rho_{red5} = 0.50, \rho_{red2} = \rho_{red6} = 0.40, \rho_{red3} = \rho_{red7} = 0.30, \rho_{red4} = \rho_{red8} = 0.20,$ are the correlation levels of the redundant features, and $\rho_{ind_m} = \rho * \rho_{red_m}$ are the correlation levels induced by ρ and ρ_{red_m} for all m=1,…,8. $\Sigma^i_{F_j,F_k}$ is the correlation matrix between the features contained in feature set j and feature set k in class i (i = 0,1; j, k = 1,2). Now, let F_1^0 be the features from feature set 1 in class 0 and F_1^1 be the features from feature set 1 in class 1 where $F_1 = F_1^0 \cup F_1^1$. Let μ_1^0 be the mean of feature set 1 in class 0 and μ_1^1 be the mean of feature set 1 in class 1. Let $F_1^0 \sim N(\mu_1^0, \Sigma^0_{F_1,F_1})$ and $F_1^1 \sim N(\mu_1^1, \Sigma^1_{F_1,F_1})$ where $\mu_1^0 = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)^T$ and $\mu_1^1 = (0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0,0,0,0,0,0,0,0)^T$. Let $F_2 = F_2^0 \cup F_2^1$ where $F_2^0 \sim N(\mu_2^0, \Sigma^0_{F_2,F_2})$ and $F_2^1 \sim N(\mu_2^1, \Sigma^1_{F_2,F_2})$ where $\mu_2^0 = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)^T$ and $\mu_2^1 = (0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0,0,0,0,0,0,0,0)^T$. Only the low

sample size was used for this experiment. Three data sets were generated and used as shown in the process flow diagram in Figure 1.

In this problem, the feature set was expanded for a more thorough examination of feature selection on a single run for one sample size (50 exemplars in each class) with one level of correlation ($\rho = 0.8$) of the same process. In this case, data was generated as described above, and discriminant loadings were calculated in the same manner as in the first feature selection problem. In this problem, the classification accuracy is defined to be the sum of the true positive values and true negative values (the sum of the correct classifications). The number of features was reduced from all 18 features for each classifier to 1 feature for each classifier. For each number of features, the classification accuracy was calculated for each fusion method.

Experimental Design

In the data generation, three data sets were generated. For all of the problems described in Table 4, the same general experimental design was followed. The only difference was in the data generation phase of the process. The first data set was used to train the individual classifiers, the linear and quadratic discriminant functions. Once the individual classifiers were trained, the second data set was used to validate the individual classifiers. Posterior probabilities were calculated from the second and third data sets for later use. The second data set posterior probabilities, in addition to being validation data for the individual classifiers, were used to train the fusion methods. The posterior probabilities from the third data set were used to validate the fusion methods. All of the plots generated are results from the third data set.

ISOC Application

The ISOC Fusion model takes a given threshold, 0.5, and determines the optimal fusion rule. Using the methodology described in the Literature Review, the optimal rule was calculated from the posterior probabilities from the second data set. After the optimal rule was calculated, the thresholds for both individual classifiers were varied together from 0.0 to 1.0, and the optimal rule was applied on the independent posterior probabilities from the third data set. The false positive values were plotted against the true positive values as the thresholds were varied; the result is six ISOC curves, one for each level of correlation. This process was replicated, and the average ISOC curve from the replications was calculated.

ROC "Within" OR Application

The ROC "Within" Fusion model takes a given rule, the "Logical OR" rule, and determines the optimal thresholds for each individual classifier. Using the methodology described in the Literature Review, the optimal threshold pairs were calculated from the posterior probabilities from the second data set. After the optimal threshold pairs were calculated, the optimal threshold pairs along with the "Logical OR" rule were applied to the independent posterior probabilities from the third data set. The false positive values were plotted against the true positive values as the thresholds were varied; the result is six ROC curves, one for each level of correlation. This process was replicated, and the average ROC curve from the replications was calculated.

PNN Application

The PNN Fusion model treats the posterior probabilities from the individual classifiers as features and outputs an overall posterior probability of an exemplar being in

a given class using the methodology described in the Literature Review. Unlike the ISOC Fusion model and the ROC "Within" Fusion model, the PNN only observes the posterior probabilities from the third data set. The PNN is trained on the first 1/3 of the data set and validated on the last 2/3 of the data set. On the validation set, the threshold was varied from 0.0 to 1.0. The false positive values were plotted against the true positive values as the thresholds were varied; the result is six ROC curves, one for each level of correlation. This process was replicated, and the average ROC curve from the replications was calculated.

One Big Network Application

The One Big Network model eliminates the individual classifiers and takes all features as inputs to a generalized regression neural network using the methodology described in the Literature Review Section (reference actual section number). The false positive values were plotted against the true positive values as the thresholds were varied; the result is six ROC curves, one for each level of correlation. This process was replicated, and the average ROC curve from the replications was calculated.

Feature Selection

In problems where each classifier observes many features, many of which are noise or redundant features, it may be beneficial to select only those features that are relevant for classification. This prevents the classifier from being confused by those features that add little to classification accuracy.

Total Probability of Misclassification

The total probability of misclassification (TPM) is calculated by summing the two error probabilities: probability of false positive and probability of false negative. This

calculation can be made for both of the classifiers used in this thesis, the linear discriminant function and the quadratic discriminant function, by using only those features that the classifier actually observes. This calculation will give a prior estimation of the errors to be observed by the classifiers. Another approach is to calculate an overall TPM, as if all the features were observed by one big classifier. This calculation will also give a prior estimation of the errors to be observed by the fusion. As the correlation between feature sets increases, less independent information is presented to the classifiers. Thus, as the correlation between feature sets increases, the TPM is expected to increase; as less independent information is presented to the classifiers, more errors are expected.

Sample Size Variation

Another experiment was designed to examine the effects of sample size. For this experiment, the sample size is defined to be the number of exemplars in each class in each data set. For each sample size, the above fusion methods were performed, and the resulting curves were generated. It is realistic that the actual amount of correlation present in the data will not be known ahead of time. It also may be realistic to assume that there is equally likely probability of observing each of the six levels of correlation. Therefore, the six ROC curves generated for each method, one for each level of correlation, were averaged into one ROC curve for each sample size. Then, the average ROC curve for each of the sample sizes was plotted to see the effects of sample size on the fusion process.

Chapter Summary

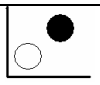
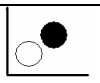
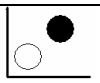
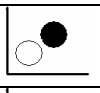
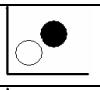
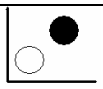
This chapter described the methodology employed throughout this thesis. First, the two types of correlation used in this thesis were presented. Then, the data generation process for each of the problems was provided. Next, the overall experimental design, as well as some general application issues for each of the four fusion methods, was given. Finally, feature selection, TPM, and sample size variation were discussed.

IV. Findings and Analysis

Introduction

After the data was generated, the four fusion methods were performed and results were generated and analyzed for each of the problems described in the data generation section. This section provides findings and analysis for each of the problems described above. Table 5 summarizes how the results are presented. All results are average results from 5 replications.

Table 5: Results Descriptions.

Problem #	Problem Name	Results Description
1 	4 Feature Case	ROC curves, N=1000 ROC curves, Across Sample Sizes
2 	8 Feature Case	ROC curves, N=1000 ROC curves, Across Sample Sizes
3 	8 Feature with Autocorrelation Case	ROC curves, N=1000 ROC curves, Across Sample Sizes
4 	8 Feature Triangle Case	ROC curves, N=1000 ROC curves, Across Sample Sizes
5 	8 Feature XOR Case	ROC curves, N=1000 ROC curves, Across Sample Sizes
6 	8 Feature XOR with Autocorrelation Case	ROC curves, N=1000 ROC curves, Across Sample Sizes
7 	20 Feature with Feature Selection Case	ROC curves, N=50 N=1000 Feature Selection vs Non-Feature Selection
8 	36 Feature with Feature Selection Case	Classification Accuracy, N=50, rho=0.8
9 	TPM Exploration	ROC Curves, N=50, N=1000 3 specific levels of correlation

Problem 1 Results: 4 Feature Case, Single Sample Size 

Features were generated according to the methodology described in the Data Generation: 4 feature without autocorrelation, and the fusion process was followed as

described in the Experimental Design section. Figure 8 shows four plots, one for each of the four fusion methods. These are average ROC curves over five replications with 1000 exemplars in each class. Each plot contains six ROC curves, one for each of the six levels of correlation. In addition, crosshairs are place at the point (0.1, 0.6) to add a point of reference common for all four plots.

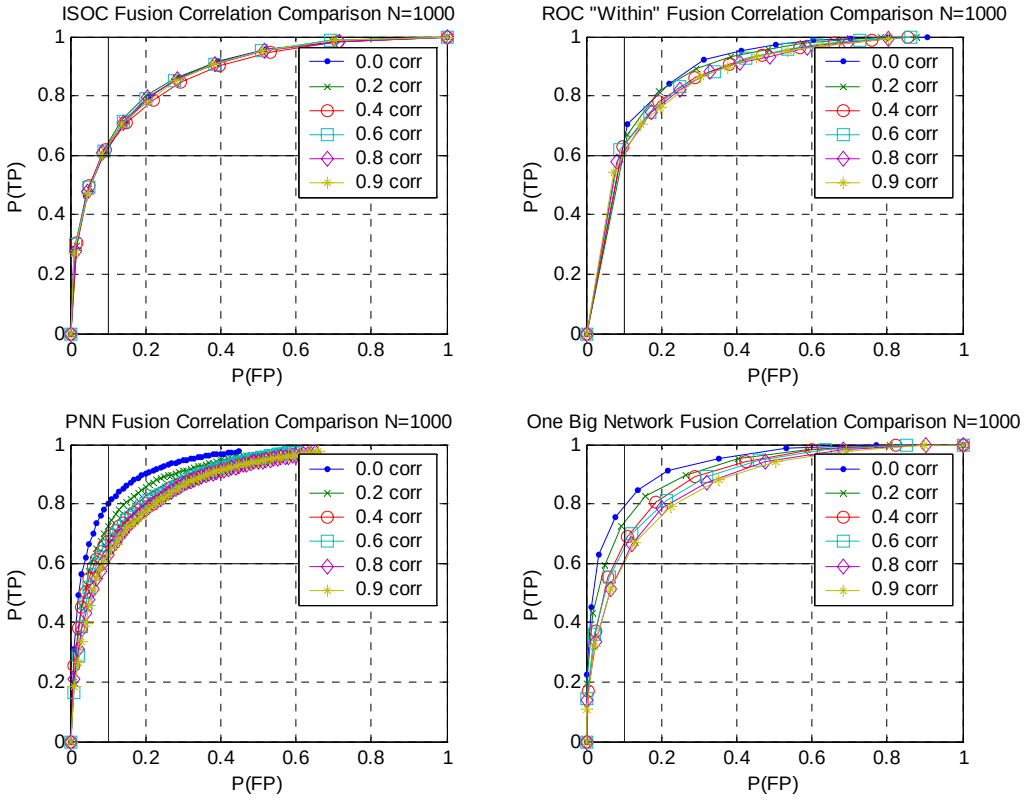


Figure 8: 4 Feature ROC Curves, N=1000.

From these plots, some meaningful conclusions can be made. First, ISOC Fusion and ROC "Within" Fusion appear to be very robust to correlation; however, they are on the low end of performance. On the other hand, the PNN is not as robust to correlation as the first two methods; that is, performance varies depending on the level of across correlation. However, the simplistic PNN performs as well as ISOC and ROC "Within"

at high levels of correlation, and it drastically outperforms them at low levels of correlation. Again, the PNN observes only half of the data that ISOC, ROC "Within", and OBN observe. The OBN approach performs comparably to the PNN.

Problem 1 Results: 4 Feature Case, Varying Sample Size 

Features were generated according to the methodology described in the Data Generation: 4 feature without autocorrelation, and the fusion process was followed as described in the Experimental Design section. This process was repeated for multiple sample sizes. Figure 9 shows four plots, one for each of the four fusion methods. They are the average ROC curves over the six levels of correlation for a given sample size. Each plot contains six ROC curves, one for each of the six sample sizes. In addition, crosshairs are place at the point (0.1, 0.6) to add a point of reference common for all four plots.

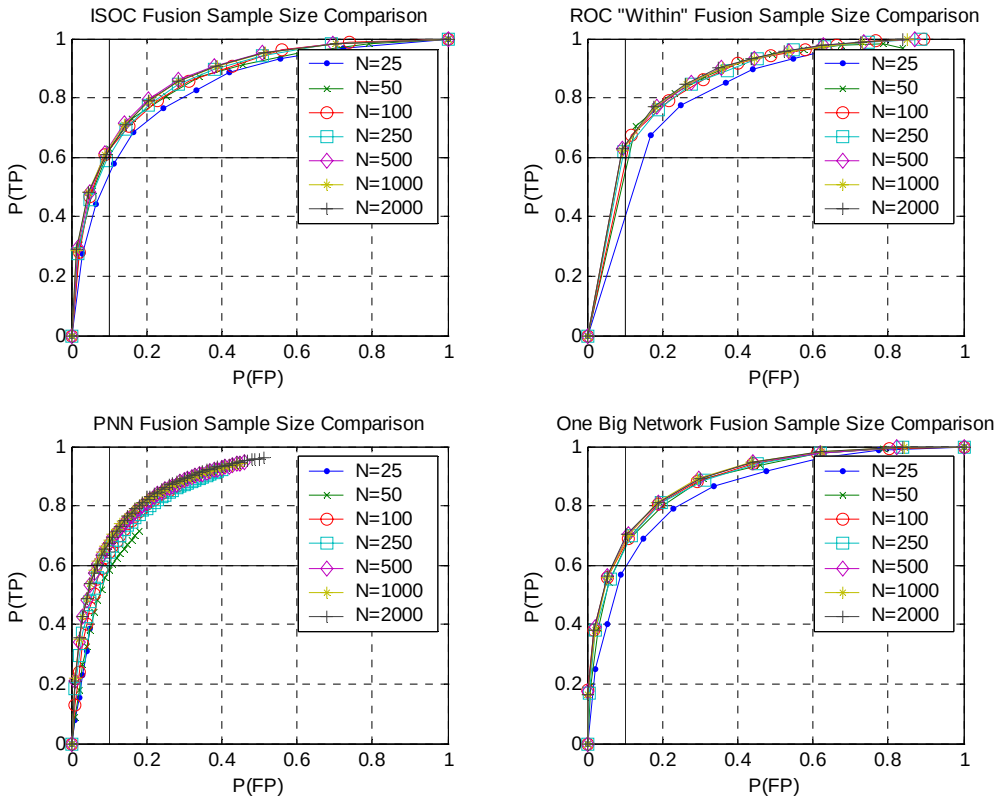


Figure 9: 4 Feature ROC Curves, Across Sample Sizes.

From these plots, it is obvious that in this simple problem, sample size is not much of a factor. For each of the 4 methods, the ROC curves for a sample size of 25 resulted in different ROC curves than the ROC curves for a sample size of 50. After a sample size of 50 is obtained, little increase in performance is gained as the sample size is increased to 2000. ISOC and OBN appear to be the most robust to sample size, and the ROC curves for the PNN appear to be better at all sample sizes.

Problem 2 Results: 8 Feature Case, Single Sample Size

Features were generated according to the methodology described in the Data Generation: 8 feature without autocorrelation, and the fusion process was followed as described in the Experimental Design section. Figure 10 shows four plots, one for each

of the four fusion methods. They are average ROC curves over five replications with 1000 exemplars in each class. Each plot contains six ROC curves, one for each of the six levels of correlation. In addition, crosshairs are place at the point (0.1, 0.4) to add a point of reference common for all four plots.

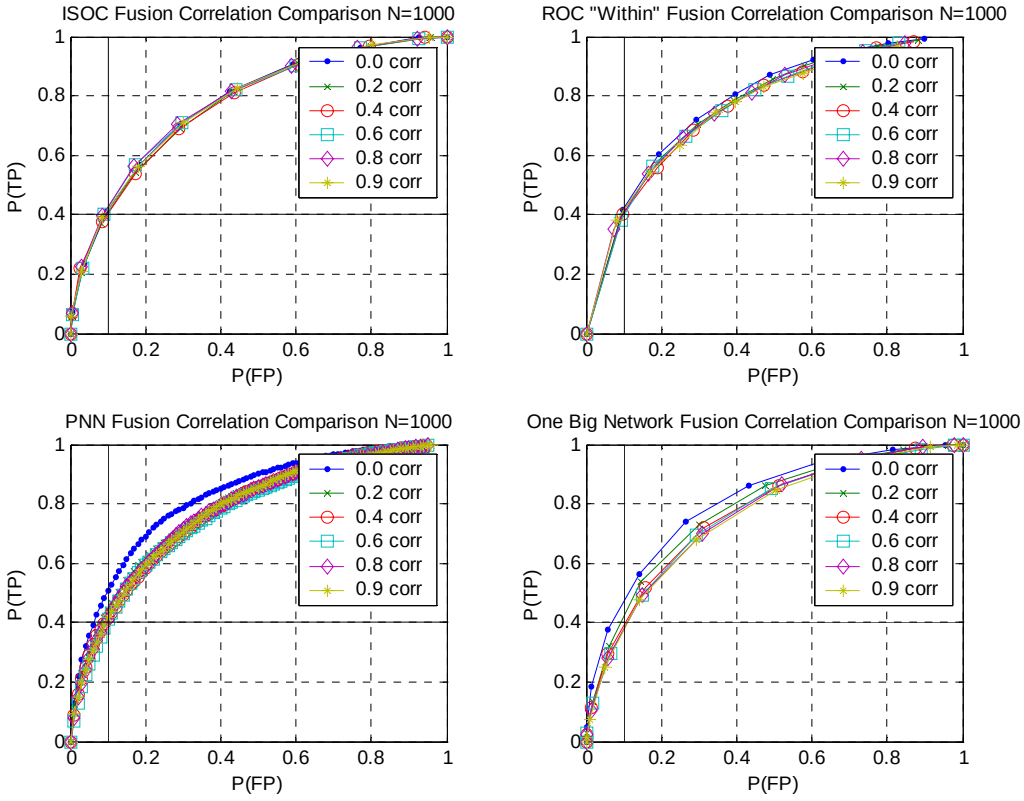


Figure 10: 8 Feature ROC Curves, N=1000.

From the plots above, many of the same conclusions can be drawn. First, as described in the data generation section, the means in this problem are closer together. That is the cause of the shift in the ROC curves; it is not the addition of the noise and redundant features. Second, as in the 4 feature problem, the ISOC and ROC "Within" Fusion methods are the most robust, but they are on the low end of performance. The

OBN is fairly robust, but it outperforms the ISOC and ROC "Within" at 0.0 level of correlation. The PNN does as well as ISOC, ROC "Within," and OBN at high levels of correlation, and it outperforms them at low levels of correlation.

Problem 2 Results: 8 Feature Case, Varying Sample Size 

Features were generated according to the methodology described in the Data Generation: 8 feature without autocorrelation, and the fusion process was followed as described in the Experimental Design section. This process was repeated for multiple sample sizes. Figure 11 shows four plots, one for each of the four fusion methods. They are the average ROC curves over the six levels of correlation for a given sample size. Each plot contains six ROC curves, one for each of the six sample sizes. In addition, crosshairs are place at the point (0.1, 0.4) to add a point of reference common for all four plots.

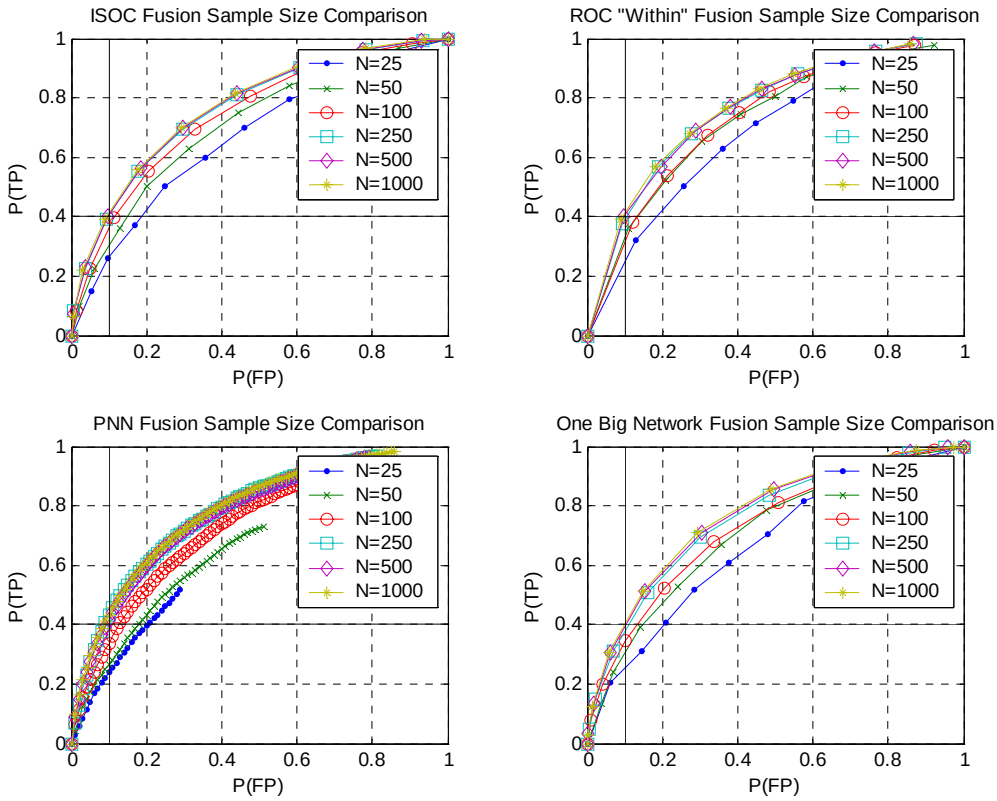


Figure 11: 8 Feature ROC Curves, Across Sample Sizes.

The sample size effect is more evident in the 8 feature case than it was in the 4 feature case. There is a fairly obvious break between the first three sample sizes and the second three sample sizes for all four fusion methods. After a sample size of 250, there is little increase in performance by increasing the sample size up to 500 or 1000. The ISOC, ROC "Within," and OBN methods seem to perform about the same for all sample sizes. The PNN does as well as the other three methods at the lower sample sizes, and it outperforms the other three methods at the higher sample sizes.

Problem 3 Results: 8 Feature with Autocorrelation Case, Single Sample Size

Features were generated according to the methodology described in the Data Generation: 8 feature with autocorrelation, and the fusion process was followed as described in the Experimental Design section. Figure 12 shows one feature over time at 0.0 level of autocorrelation, and Figure 13 shows one feature over time at 0.9 level of autocorrelation.

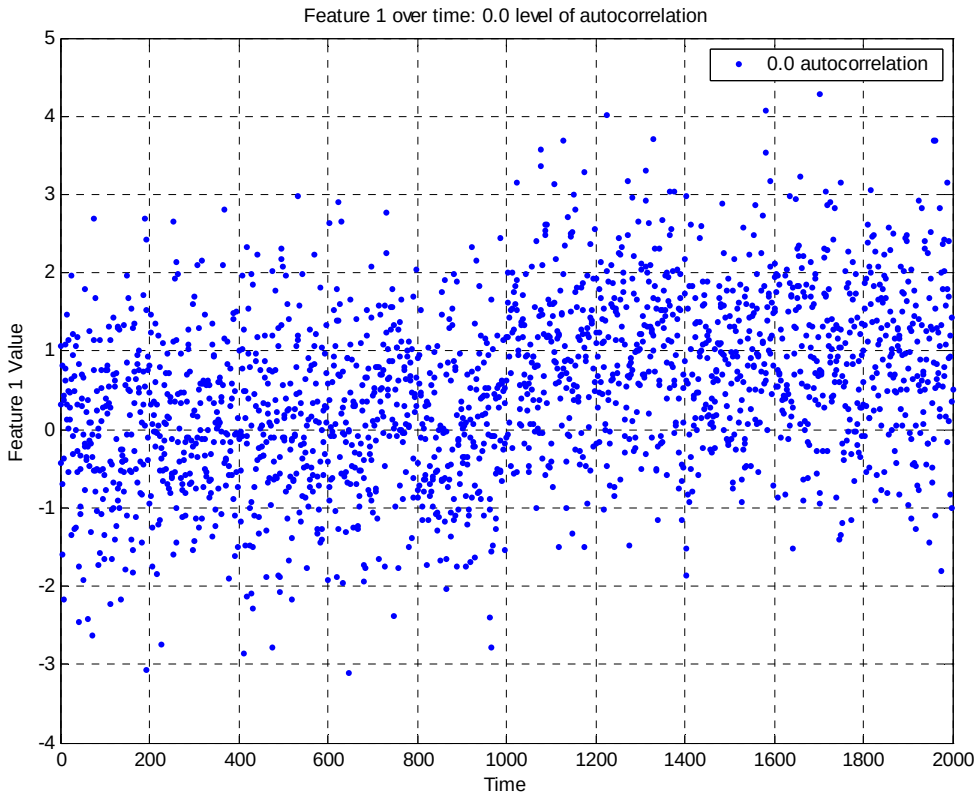


Figure 12: Feature 1 over Time: 0.0 Level of Autocorrelation.

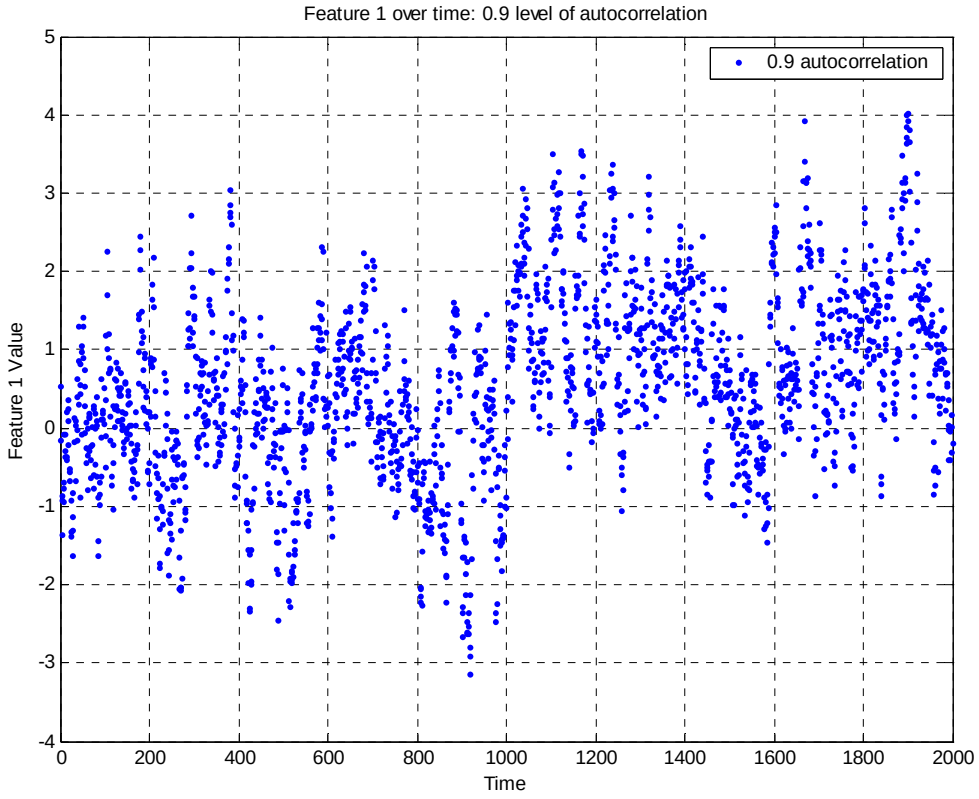


Figure 13: Feature 1 over Time: 0.9 Level of Autocorrelation.

Figure 12 shows that, as expected, feature 1 is independent over time; this is evident because there is no pattern in the data over time. Figure 13 shows that autocorrelation is present; this is evident because there is a definite pattern over time. Thus, the appropriate levels of autocorrelation are present.

Figure 14 shows four plots, one for each of the four fusion methods. They are the values of true positive for a false positive value of 0.1 on the average ROC curves over five replications with 1000 exemplars in each class. The true positive value is plotted against the level of across correlation. Each plot contains three lines, one for each of the three levels of autocorrelation.

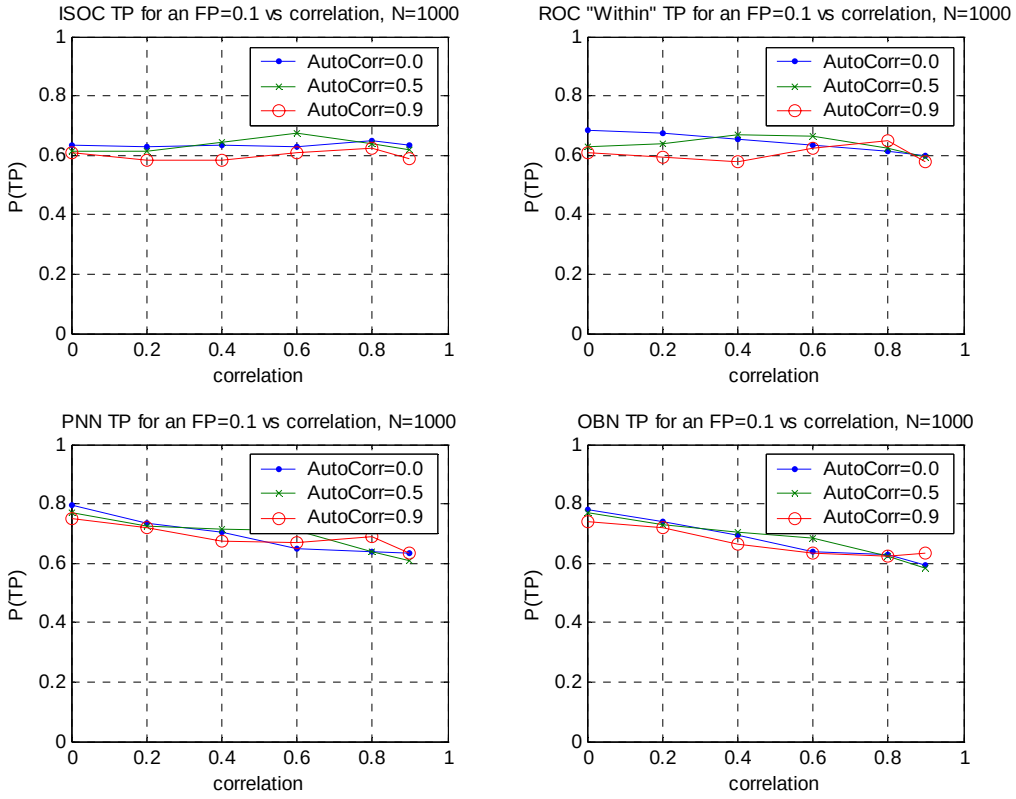


Figure 14: 8 Feature with Autocorrelation Case, N=1000.

Figure 14 shows that there is an effect of autocorrelation; although, it is not dramatic at this high sample size. It also seems to make a more significant difference in those fusion methods that assume independence of the classifiers, ISOC and ROC "Within." There is less of a difference in those fusion methods that make no such assumption about the classifiers, PNN and OBN. Once again, at this sample size, the OBN and PNN fusion far exceed the performance of ISOC and ROC "Within" at low levels of correlation and perform at least as well at high levels of correlation.

Problem 3 Results: 8 Feature with Autocorrelation Case, Across Sample Sizes

Features were generated according to the methodology described in the Data Generation: 8 feature with autocorrelation, and the fusion process was followed as described in the Experimental Design section. This process was repeated for multiple sample sizes. Figure 15 shows four plots, one for each of the four fusion methods, They are the true positive values for a false positive value of 0.1 on the average ROC curves over the six levels of correlation for a given sample size. Sample size is varied from 50 exemplars in each class to 1000 exemplars in each class. Each plot contains three lines, one for each of the three levels of autocorrelation.

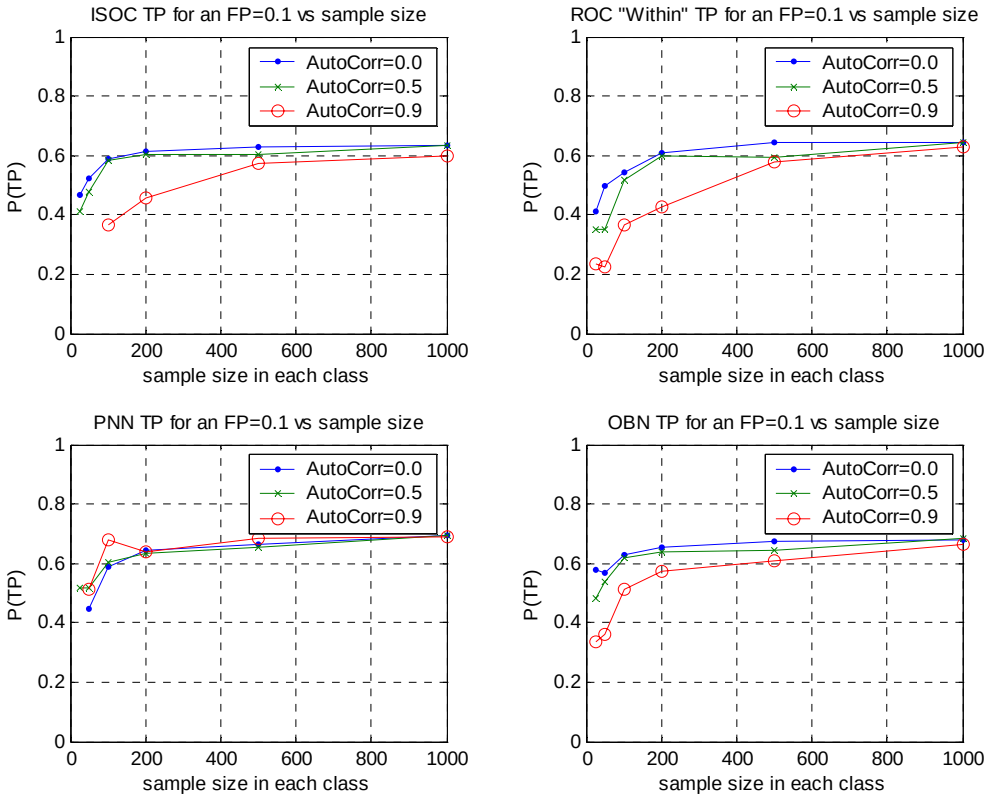


Figure 15: 8 Feature with Autocorrelation Case, Across Sample Sizes.

Figure 15 shows that the PNN is fairly robust to the autocorrelation while the other three methods are more affected by the presence of autocorrelation. This is especially true as the level of autocorrelation reaches 0.9. That is, there is little difference between 0.0 level of autocorrelation and 0.5 level of autocorrelation for any of the three methods; however, there is a difference between 0.5 level of autocorrelation and 0.9 level of autocorrelation. This degradation in performance is more dramatic as the sample size decreases. At high sample size, there is little difference between 0.0 level of correlation and 0.9 level of correlation for any of the three methods. By the time the sample size drops to 500 or 200, there is significant degradation in performance for all methods except the PNN. In conclusion, the PNN and OBN outperform the other two methods regardless of the level of autocorrelation or the sample size.

Problem 3 Results: 8 Feature with Autocorrelation Case, An ANOVA Approach

This same set of data can also be examined using an ANOVA approach. Consider a three factor design where the three factors are Level of Autocorrelation, Level of Across Correlation, and Sample Size. Level of Autocorrelation has three levels, 0.0, 0.5, and 0.9. Level of Across Correlation has six levels, 0.0, 0.2, 0.4, 0.6, 0.8, and 0.9. Sample Size has five levels, 50, 100, 250, 500, and 1000 exemplars in each class. A full factorial design consists of all possible combinations of these three factors. Each design point was replicated five times; there were a total of 450 runs for each of the four methods. Table 6 summarizes the results of each ANOVA.

Table 6: Summary of ANOVA Results.

Method	ISOC	ROC "Within"	PNN	OBN
R^2	0.406	0.371	0.372	0.383
Adjusted R^2	0.259	0.215	0.216	0.230
Mean Response	0.572	0.532	0.623	0.617
Root MSE	0.134	0.151	0.154	0.130
Factors Considered and corresponding p-values				
Autocorrelation	<.0001	<.0001	0.1639	<.0001
Sample Size	<.0001	<.0001	<.0001	<.0001
Auto*Sample	<.0001	0.1431	0.5337	0.6097
Correlation	0.8973	0.2866	<.0001	<.0001
Auto*Corr	0.3613	0.8030	0.7965	0.5096
Corr*Sample	0.6699	0.2815	0.0801	0.3318
Auto*Corr*Sample	0.8428	0.9243	0.9997	0.9995

These results seem to confirm statistically the results that were already shown graphically. ISOC and ROC "Within" are very robust to across correlation, but they are not robust to autocorrelation or sample size. In other words, changing the autocorrelation level and sample size level will have an impact on the performance of these two types of fusion. The OBN is not robust to any of the three factors. The PNN is robust to autocorrelation level, but the PNN performance will change as the level of autocorrelation and sample size change. The good news for the PNN and OBN is that although their performance is less robust to across correlation, they also do as well or better than the other two methods. This is also apparent in the Mean Response for each method. PNN has the highest mean response, and OBN has the second highest mean response. ISOC has the third highest mean response, and ROC "Within" has the lowest mean response. It is also worth noting that none of the R^2 or Adjusted R^2 are particularly good. This means that only approximately 20% of the variation in the data can be explained by the particular models.

Validation of the assumptions, normal errors and constant variance, via residual analysis was done for each of the four methods. All four followed the same pattern so only those for ISOC are shown. Figure 16 shows a histogram of the residuals; Figure 17 shows the residuals vs Row Number.

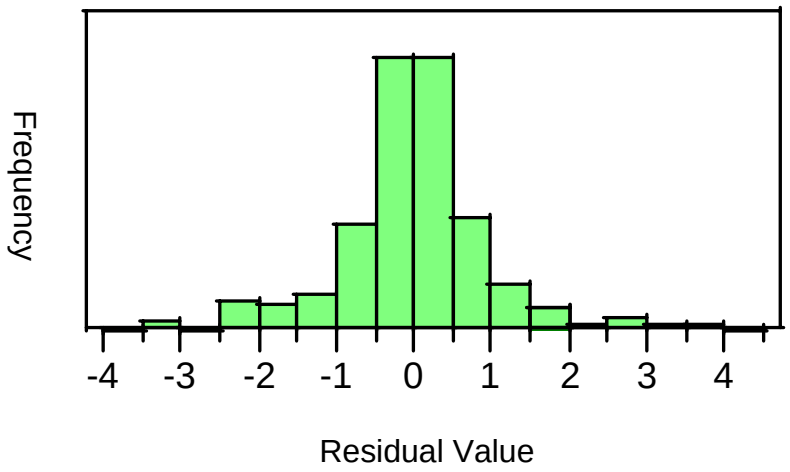


Figure 16: ISOC Histogram of Residuals.

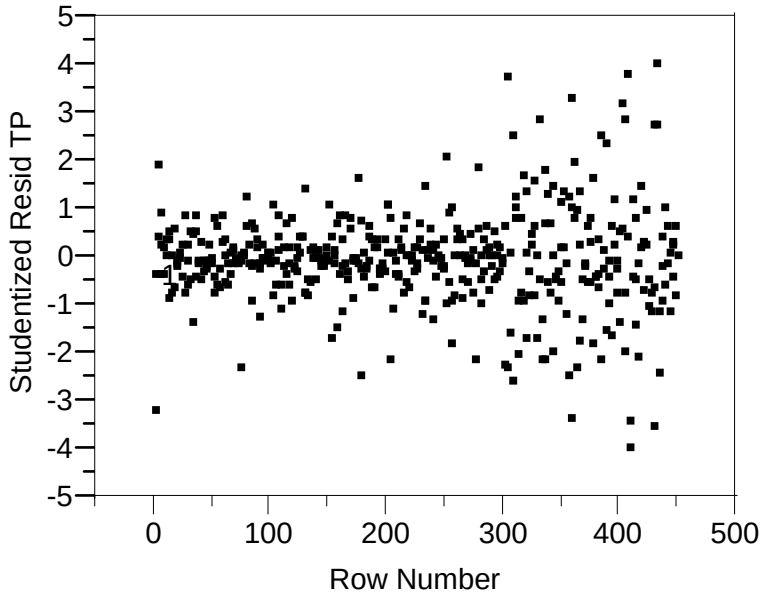


Figure 17: ISOC Residual TP Probability vs Row Number.

Figure 16 shows that the errors are approximately normally distributed so the normal error assumption holds. Figure 17 shows that there is a violation of the constant variance assumption. This means that the variance of the residuals is not constant over the entire process. This plot can be divided into three parts: the first 150 rows correspond to 0.0 level of autocorrelation, the second 150 rows correspond to 0.5 level of autocorrelation, and the final 150 rows correspond to 0.9 level of autocorrelation. The first 300 rows seem to have constant variance; that is, for all the responses for 0.0 level of autocorrelation and 0.5 level of autocorrelation, the variance seems to stay the same regardless of the level of across correlation or sample size. The variance of the residuals seems to explode right around row 300. This is the start of the 0.9 level of autocorrelation. At this point, there is a much wider variance in responses than for the first two levels of autocorrelation. Figure 18 is a similar plot, but the rows are in slightly

different order. Now, the data is first sorted by autocorrelation in ascending order and then by sample size in descending order.

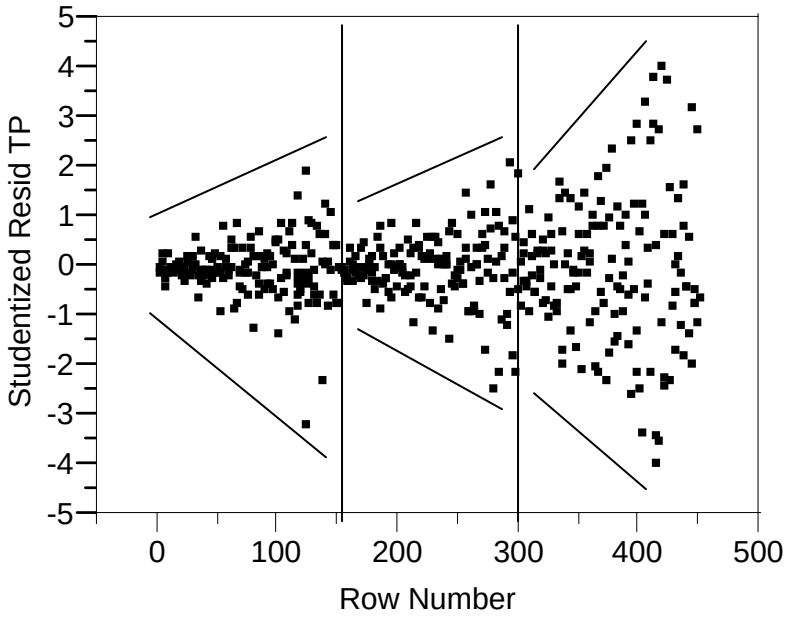


Figure 18: ISOC Residuals vs Row Number, Resorted

Figure 18 shows even more discernable pattern of heteroskedasticity. Within each group of 150 rows, there is an even more evident pattern. As the sample size decreases, the variance within a level of autocorrelation increases. Thus, one can expect the variance to increase not only as the autocorrelation level increases, but also as the sample size decreases.

Since there is so much variability in the data, another approach is to average the five replications for each design point and use the average TP value as the response. Table 7 shows the results for this analysis.

Table 7: Summary of ANOVA Results, Averaged.

Method	ISOC	ROC "Within"	PNN	OBN
R^2	0.871	0.894	0.926	0.906
Adjusted R^2	0.705	0.764	0.835	0.791
Mean Response	0.557	0.524	0.623	0.604
Root MSE	0.060	0.064	0.043	0.051
Factors Considered and corresponding p-values				
Autocorrelation	<.0001	<.0001	0.0160	<.0001
Sample Size	<.0001	<.0001	<.0001	<.0001
Auto*Sample	0.0005	0.0040	0.0459	0.0053
Correlation	0.7204	0.8582	<.0001	<.0001
Auto*Corr	0.2231	0.0444	0.1519	0.2656
Corr*Sample	0.4223	0.1227	0.0002	0.1448

The same trends hold in this analysis in terms of the most significant factors for a given method. Also, the mean responses hold the same ranking. That is, PNN has the highest mean response, and OBN has the second highest mean response. ROC "Within" has the lowest mean response, and ISOC has the second lowest mean response. This analysis does reduce the root mean square error since there is less inherent variability in this process. Finally, the R^2 and Adjusted R^2 values for all four methods are much higher than they were in the previous analysis. Again, since there is much more variability in this data, the model does a much better job of explaining the variation in the data.

Validation of the assumptions, normal errors and constant variance, via residual analysis was also done for each of the four methods in this analysis. All four followed the same pattern so only those for ISOC are shown. Figure 19 shows a histogram of the residuals; Figure 20 shows the residuals over time.

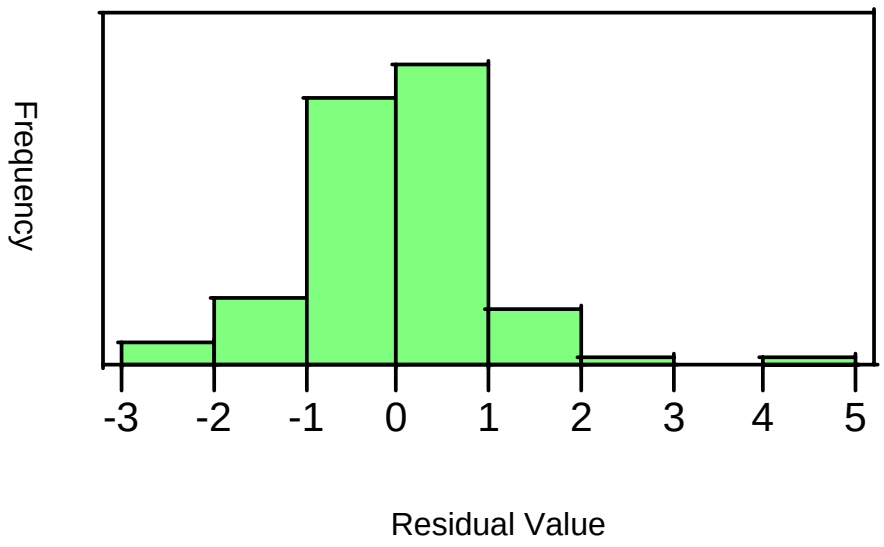


Figure 19: ISOC Histogram of Average Residuals

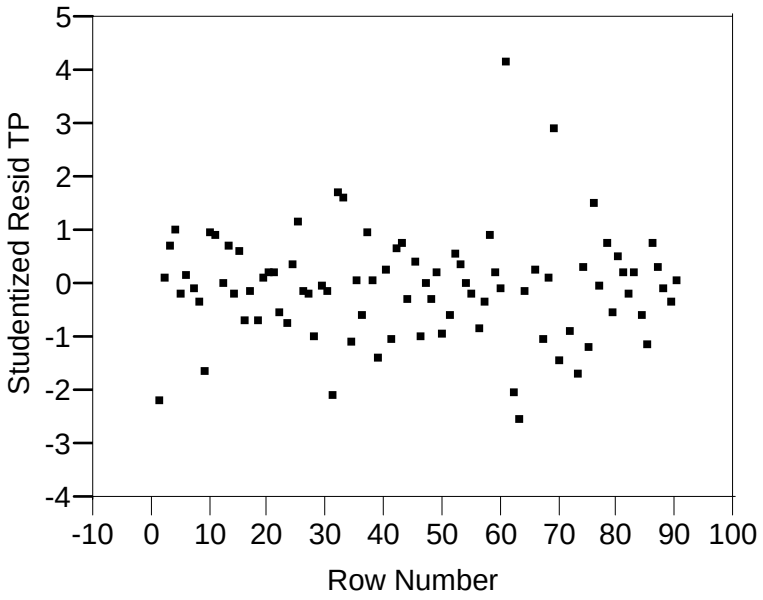


Figure 20: ISOC Residual TP Probability vs Row Number.

These residual plots show resolution to the violation of assumptions above.

Figure 19 shows that the residuals are approximately normally distributed, and Figure 20

shows that there is approximately a constant variance over the entire process. This means that the data in this new analysis does not violate either of the two assumptions.

Problem 4 Results: 8 Feature Triangle Case, Single Sample Size

Features were generated according to the methodology described in the Data Generation: 8 feature Triangle Problem, and the fusion process was followed as described in the Experimental Design section. Figure 21 shows four plots, one for each of the four fusion methods. They are average ROC curves over five replications with 1000 exemplars in each class. Each plot contains six ROC curves, one for each of the six levels of correlation. In addition, crosshairs are place at the point (0.1, 0.4) to add a point of reference common for all four plots.

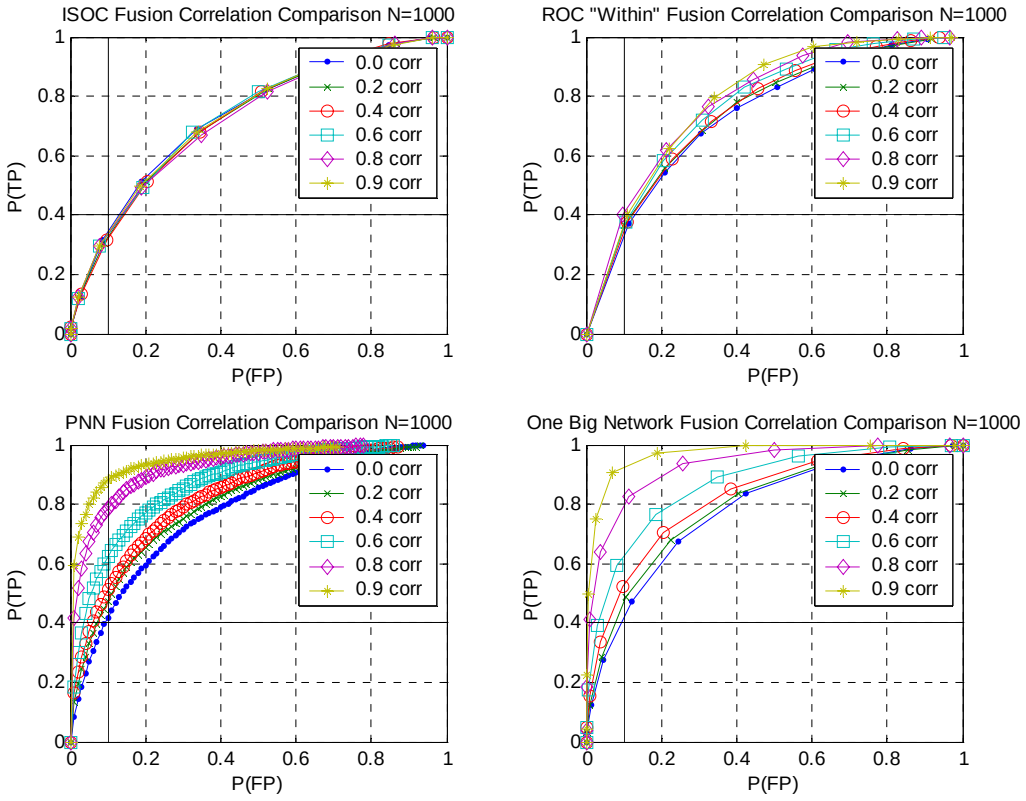


Figure 21: 8 Feature Triangle ROC Curves, N=1000.

There is a wide difference between the four methods in this problem. First of all, ISOC is the most robust to correlation, and ROC "Within" is second most robust to correlation. Although, they are the most robust to correlation, they are both on the very low end of performance. On the other hand, the PNN and OBN are not very robust to correlation, but they outperform the other two methods at all levels of correlation. They far outperform the other two methods at high levels of correlation and slightly outperform at low levels of correlation. In general, in the simpler problems, the higher the level of correlation results in a lower the level of performance. In this problem, there is an inverse effect; this is particularly evident in the PNN and OBN results. The higher level of correlation results in a higher level of performance.

At first glance, this result seems highly counter-intuitive, but a look at the geometry of the problem provides valuable insight into the results for the PNN. Figure 22 shows a plot of the posterior probabilities from the linear classifier vs the posterior probabilities from the quadratic classifier for the 0.0 correlation level. In essence, this is the feature space of the PNN.

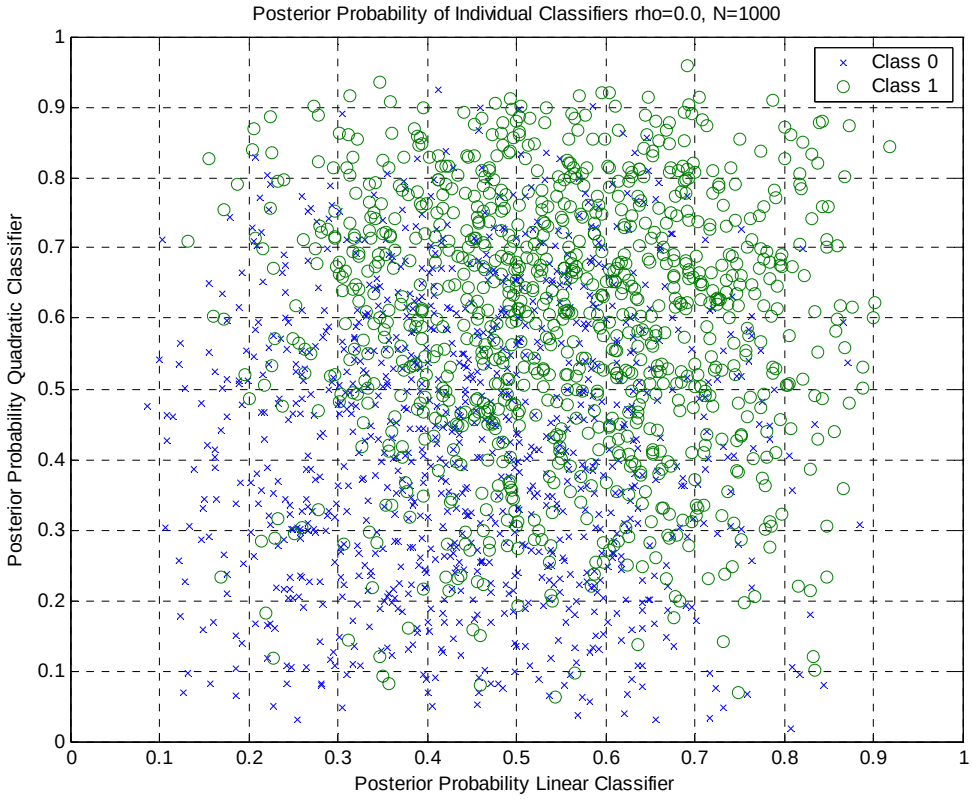


Figure 22: PNN Feature Space Plot, 0.0 Correlation, N=1000.

Figure 22 shows very little separation between the two classes; thus, it is difficult for the PNN to distinguish between the two classes. However, this plot is much different when the correlation is 0.9. Figure 23 shows a plot of the posterior probabilities from the linear classifier vs the posterior probabilities from the quadratic classifier for the 0.9 correlation level.

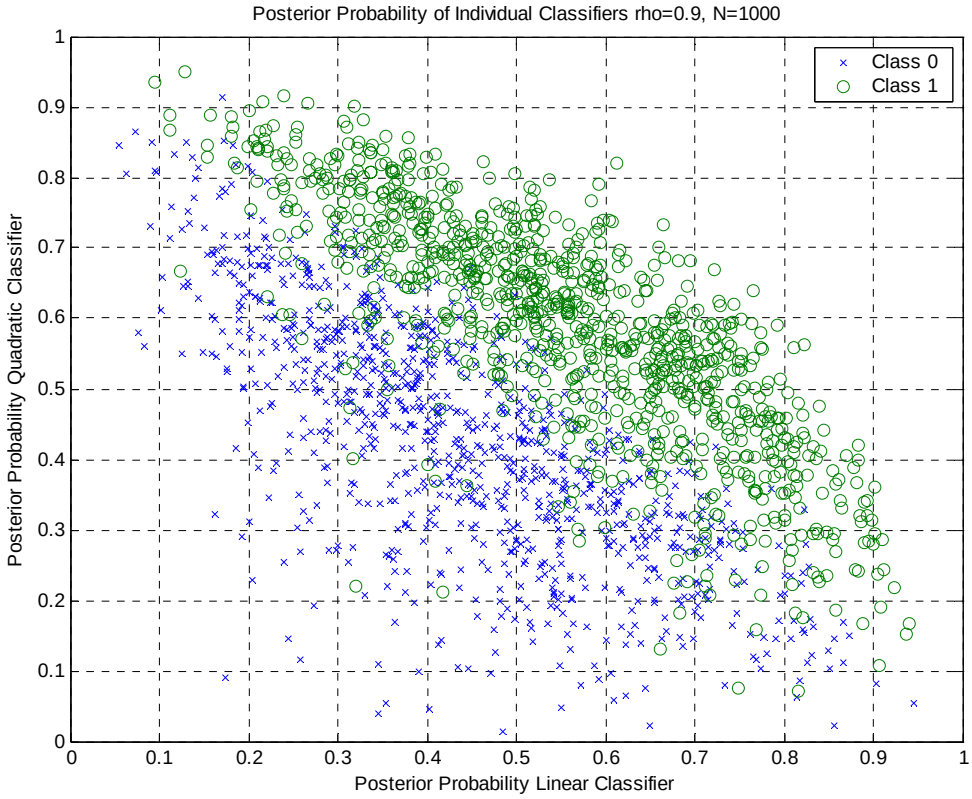


Figure 23: PNN Feature Space Plot, 0.9 Correlation, N=1000.

As is obvious from Figure 23, adding the correlation significantly alters the geometry of the problem posed to the PNN. Now, the PNN can very easily solve the problem since the classes are much more distinguishable. This explains why, in this problem, the PNN has increased performance as the correlation level increases.

The OBN results also seem highly counter-intuitive, but as was the case with the PNN, a look at the geometry of the OBN problem provides valuable insight into the results. Figure 24 shows the feature space of feature 1 and feature 2 at 0.0 level of correlation. Figure 25 shows the feature space of feature 1 and feature 4 at 0.0 level of correlation.

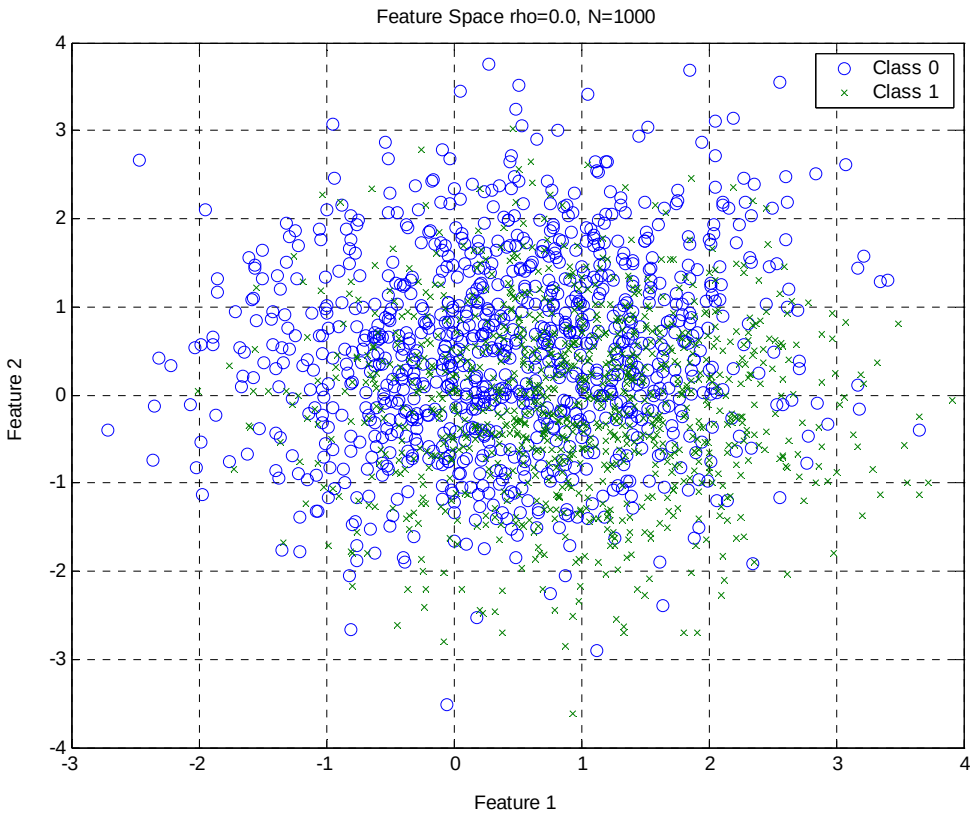


Figure 24: Feature Space of Feature 1 and Feature 2, 0.0 Correlation, N=1000.

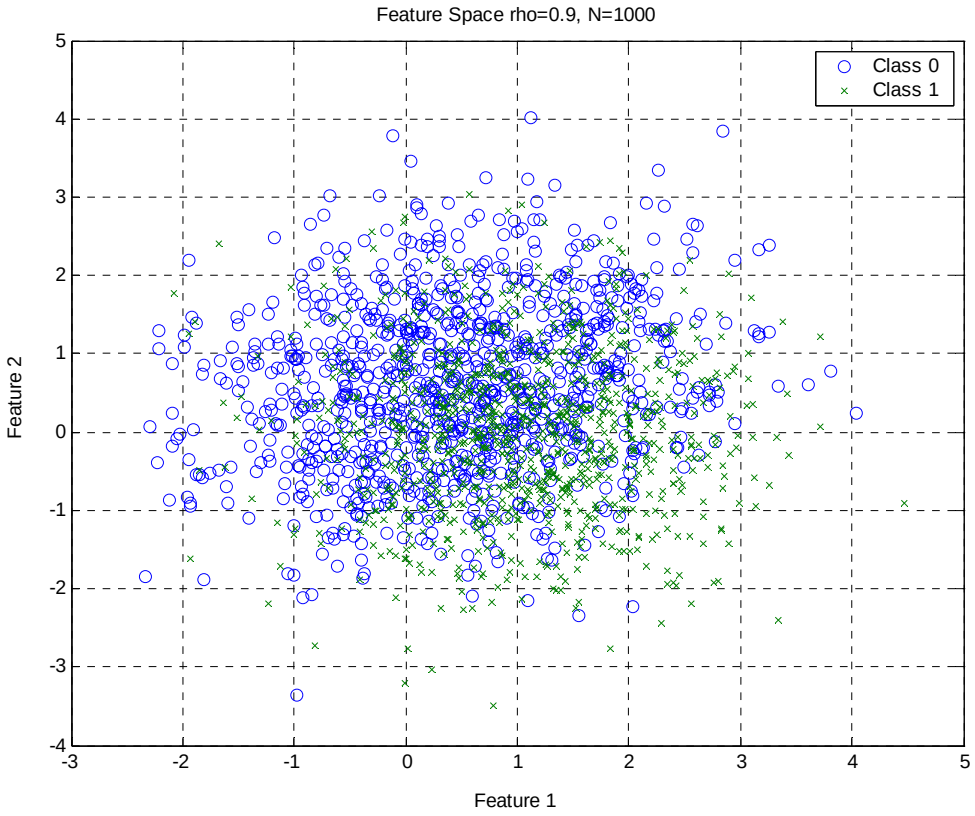


Figure 25: Feature Space of Feature 1 and Feature 2, 0.9 Correlation, N=1000.

Neither of these plots is unexpected. Since Feature 1 and Feature 2 are always independent, the correlation level does not change the geometry of the plot. Figure 24 shows very little separation between the two classes in the feature space of feature 1 and feature 2 at 0.0 level of correlation, and Figure 25 shows very little separation between the two classes in the feature space of feature 1 and feature 2 at 0.9 level of correlation. Thus, it is difficult for the OBN to distinguish between the two classes in this dimension regardless of correlation.

Figure 26 shows the feature space of feature 1 and feature 4 at 0.0 level of correlation. Figure 27 shows the feature space of feature 1 and feature 4 at 0.9 level of correlation.

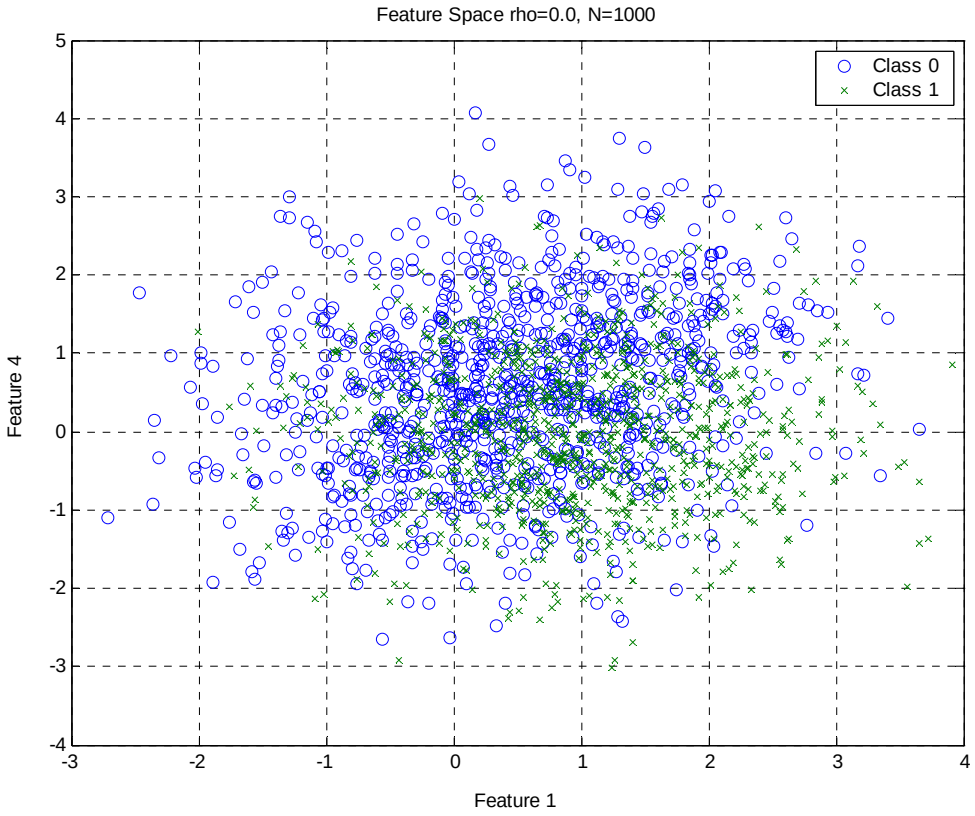


Figure 26: Feature Space of Feature 1 and Feature 4, 0.0 Correlation, N=1000.

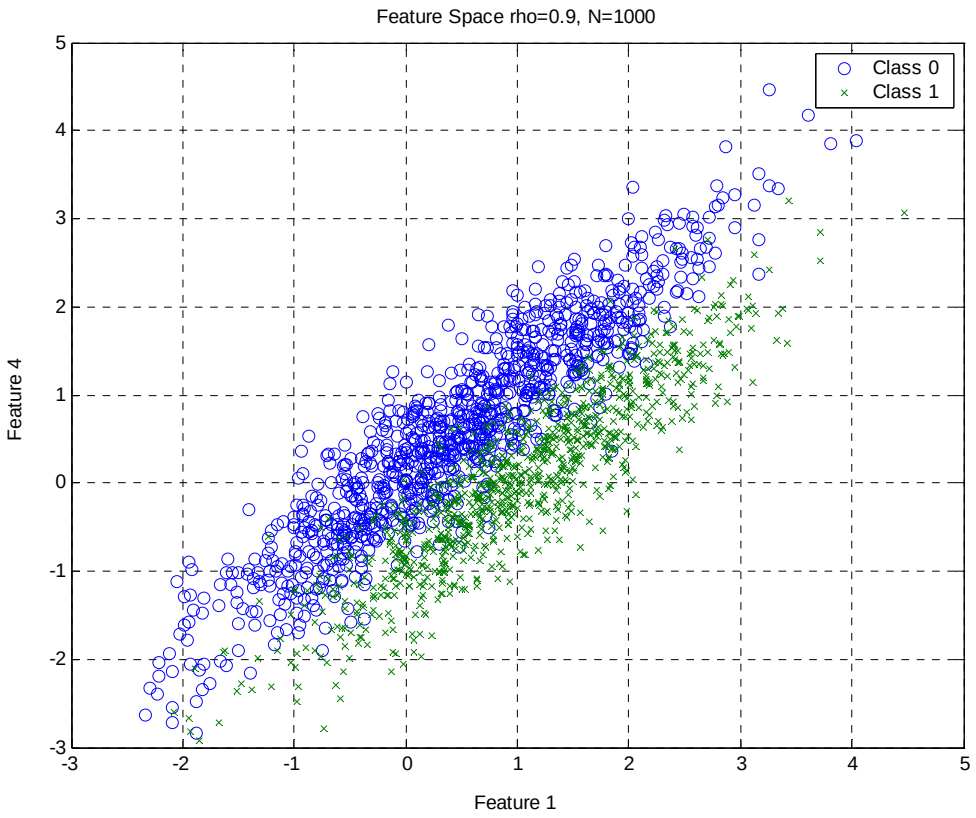


Figure 27: Feature Space of Feature 1 and Feature 4, 0.9 Correlation, N=1000.

Since there is no correlation between features 1 and 4 in Figure 26, there is still little separation between the classes. At a high level of correlation, the shape of the feature space in this dimension is significantly changed such that the OBN can easily tell the difference between the two classes. This change in geometry explains why the OBN performance increases as the level of correlation increases. As the problem is complicated slightly, the ISOC and ROC "Within" methods continue to diminish in performance while the OBN and PNN methods continue to outperform at some levels of correlation.

Problem 4 Results: 8 Feature Triangle Case, Varying Sample Size

Features were generated according to the methodology described in the Data Generation: 8 feature Triangle Problem, and the fusion process was followed as described in the Experimental Design section. This process was repeated for multiple sample sizes. Figure 28 shows four plots, one for each of the four fusion methods. They are the average ROC curves over the six levels of correlation for a given sample size. Each plot contains six ROC curves, one for each of the six sample sizes. In addition, crosshairs are place at the point (0.1, 0.4) to add a point of reference common for all four plots.

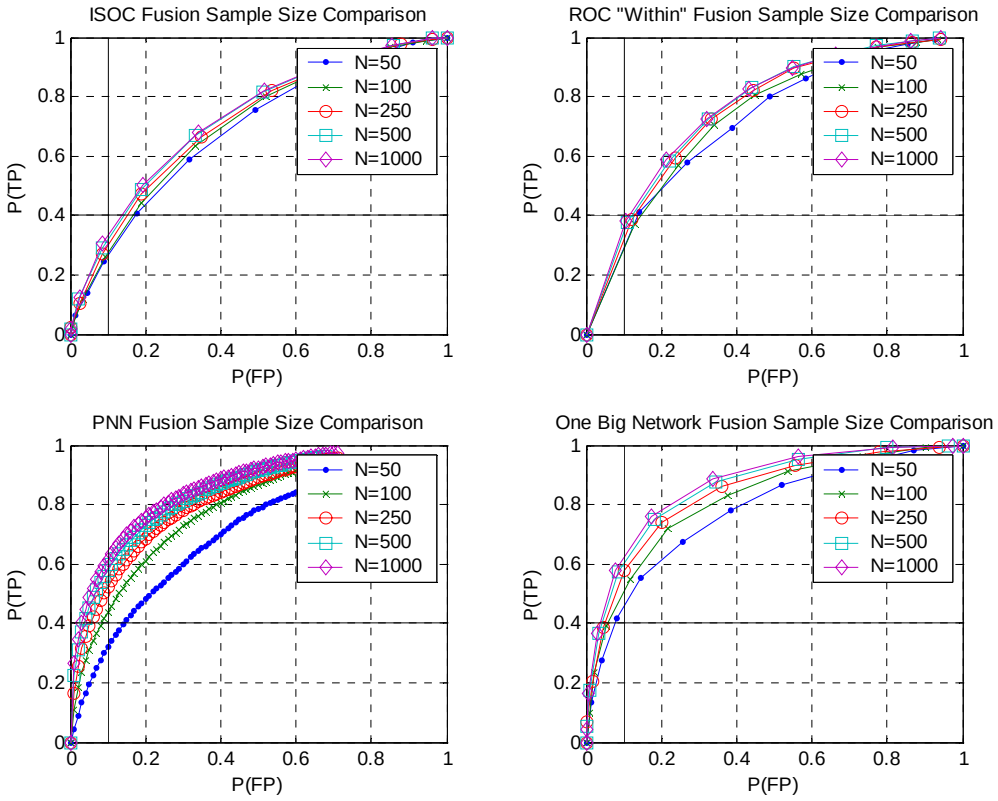


Figure 28: 8 Feature Triangle ROC Curves, Across Sample Sizes.

As the problem becomes more complicated, more samples are needed. The first three sample sizes are fairly well separated in all four methods while there is little difference between N=500 and N=1000 for any of the four methods. This means that once the sample size is 500, there is little to be gained by increasing the sample size in this problem. Again, as the problem is slightly more complicated, the ISOC and ROC "Within" methods continue to diminish in performance while the OBN and PNN methods continue to outperform at some levels of correlation.

Problem 5 Results: 8 Feature XOR Case, Single Sample Size 

Features were generated according to the methodology described in the Data Generation: 8 feature XOR Problem, and the fusion process was followed as described in the Experimental Design section. Figure 29 shows four plots, one for each of the four fusion methods. They are average ROC curves over five replications with 1000 exemplars in each class. Each plot contains six ROC curves, one for each of the six levels of correlation. In addition, crosshairs are place at the point (0.1, 0.2) to add a point of reference common for all four plots.

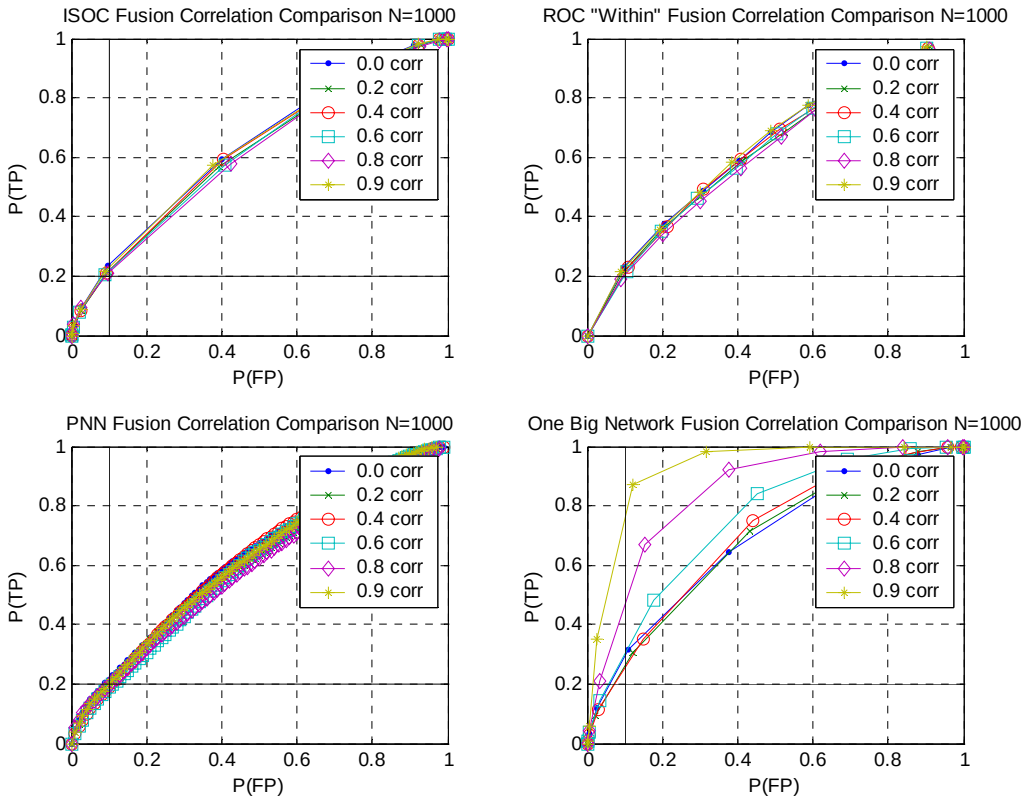


Figure 29: 8 Feature XOR ROC Curves, N=1000.

A linear discriminant function can not adequately solve the XOR problem so the linear classifier is not a good classifier. With adequate separation of the classes, the quadratic discriminant function could solve the XOR problem, but in this problem without adequate separation, the quadratic classifier is not a good classifier either. Thus, both the ISOC and ROC "Within" can not improve upon fusing two bad classifiers. Now that the problem has become too complicated for either the linear or quadratic classifier to adequately solve, the posterior probabilities from each of the classifiers do not offer much more information than the binary predicted classes. Thus, the PNN does not improve upon either of ROC "Within" or ISOC Fusion methods. On the other hand, since the OBN eliminates the bad classifiers altogether, it is able to mildly outperform the

other three methods at low levels of correlation, and it is able to greatly outperform the other three methods at high levels of correlation. This improvement as a result of increasing correlation is due to similar geometric effects as were shown in previous sections.

Interestingly, even when the classes are further separated, the PNN only does as well as the ISOC and ROC "Within." There is never a huge increase in performance at any levels of correlation as there was in the simpler problems. To show this, the above problem was rerun with further separation in the classes for a single sample size. Figure 30 shows four plots, one for each of the four fusion methods where there is further separation in the classes.

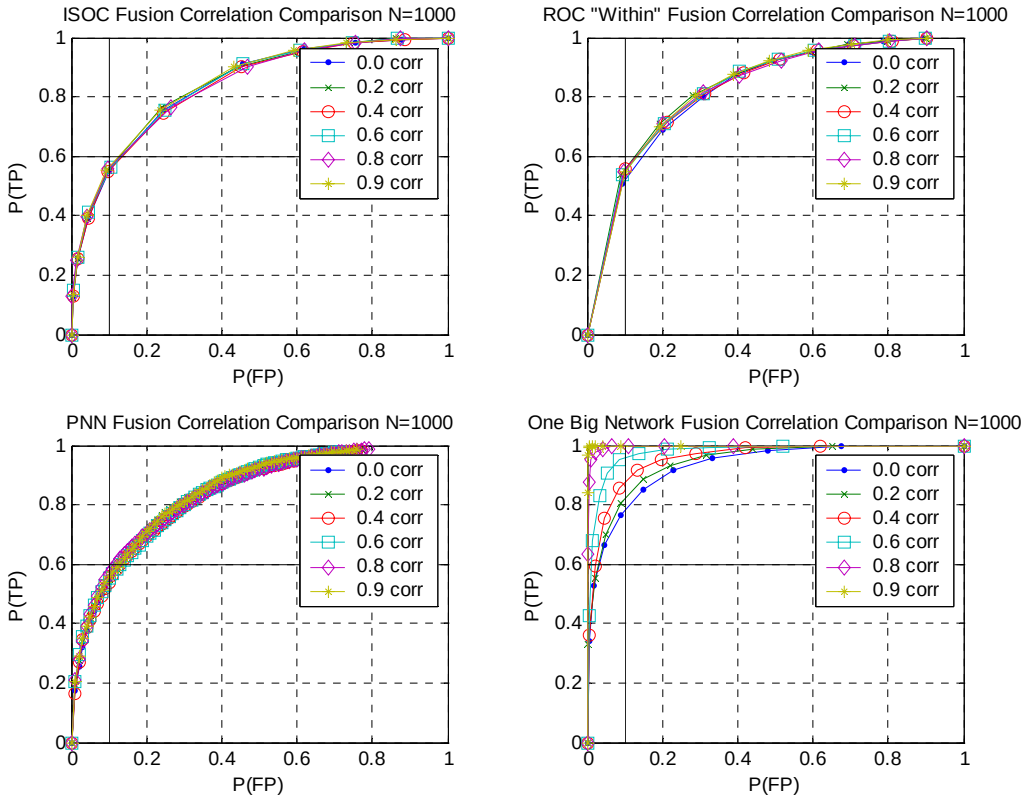


Figure 30: 8 Feature XOR ROC Curves with More Separation, N=1000.

Again, the PNN performs only as well as the ISOC and ROC "Within" fusion methods. This is explained with further investigation. Since the linear classifier is a bad classifier in this XOR problem, the posterior probabilities from the linear classifier are also bad. Thus, the PNN fusion method only uses the information from the quadratic classifier. Figure 31 shows the individual classifier average ROC curves over 5 replications for 0.0 correlation and Figure 32 shows the individual classifier average ROC curves over 5 replications for 0.9 correlation.

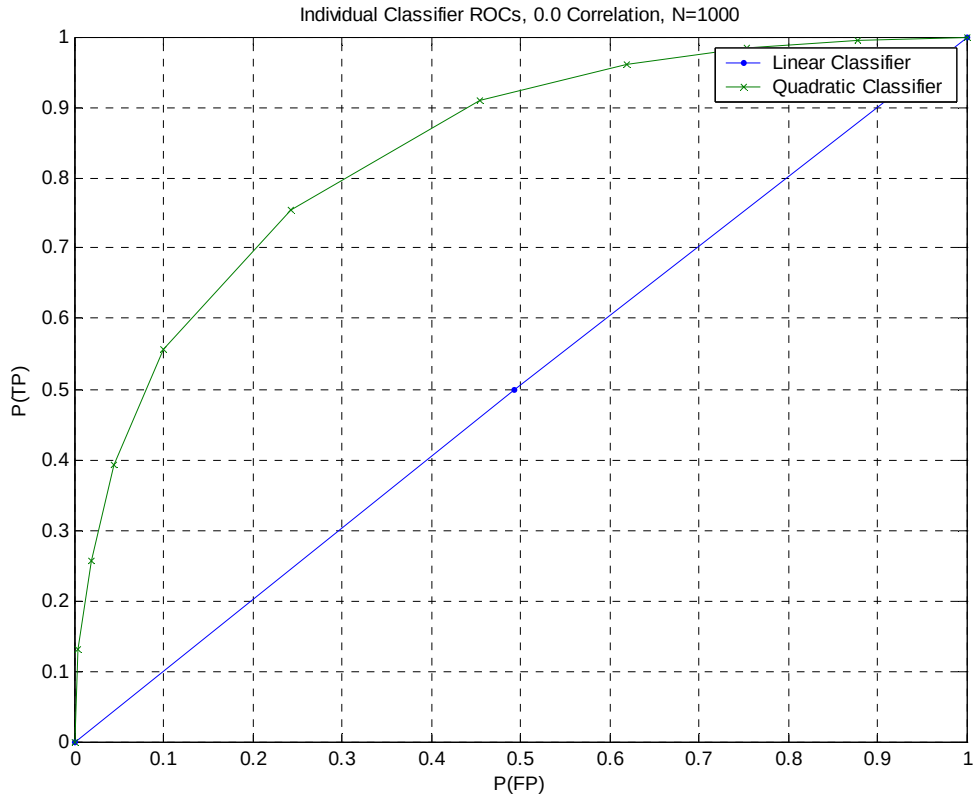


Figure 31: Individual Classifier ROC Curves for 0.0 Correlation.

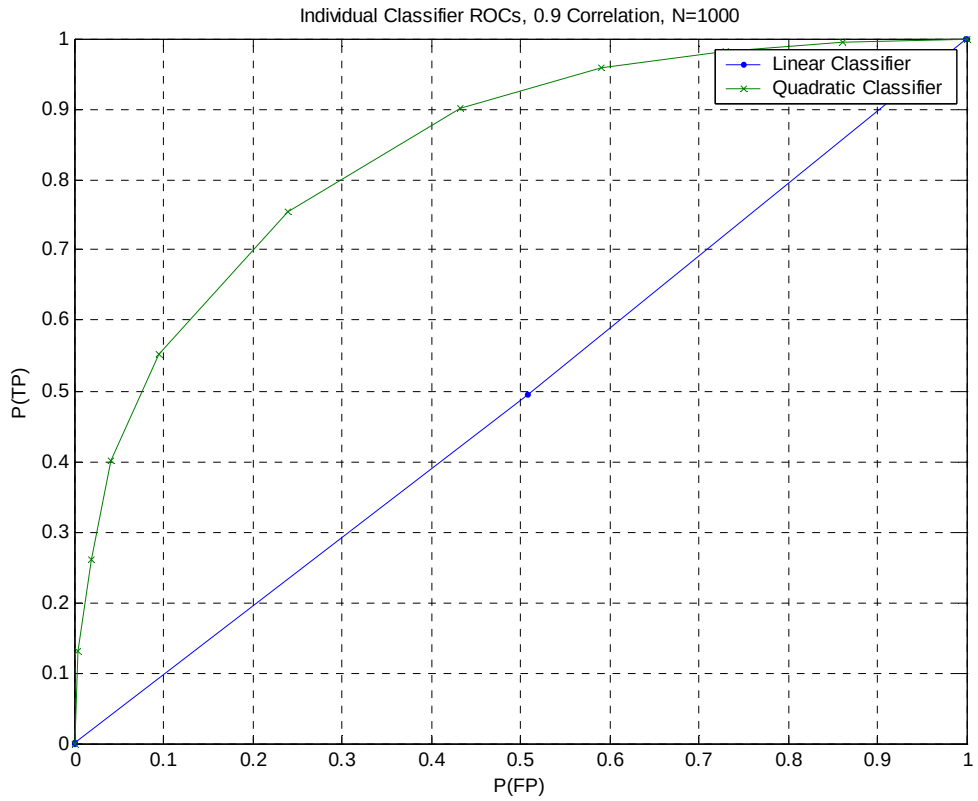


Figure 32: Individual Classifier ROC Curves for 0.9 Correlation.

Since the quadratic classifier observes independent information, regardless of the across correlation, the performance of this classifier does not change as the level of correlation changes. Also, the performance of the PNN fusion is almost identical to the performance of the quadratic classifier. This explains the difference in performance in the PNN fusion between this type of problem and the simpler type of problem. The OBN continues to outperform the other methods in this type of problem because of the same geometric explanation in the previous problem.

Problem 5 Results: 8 Feature XOR Case, Varying Sample Size 

Features were generated according to the methodology described in the Data Generation: 8 feature XOR Problem, and the fusion process was followed as described in

the Experimental Design section. This process was repeated for multiple sample sizes. Figure 33 shows four plots, one for each of the four fusion methods. They are the average ROC curves over the six levels of correlation for a given sample size. Each plot contains six ROC curves, one for each of the six sample sizes. In addition, crosshairs are place at the point (0.1, 0.2) to add a point of reference common for all four plots.

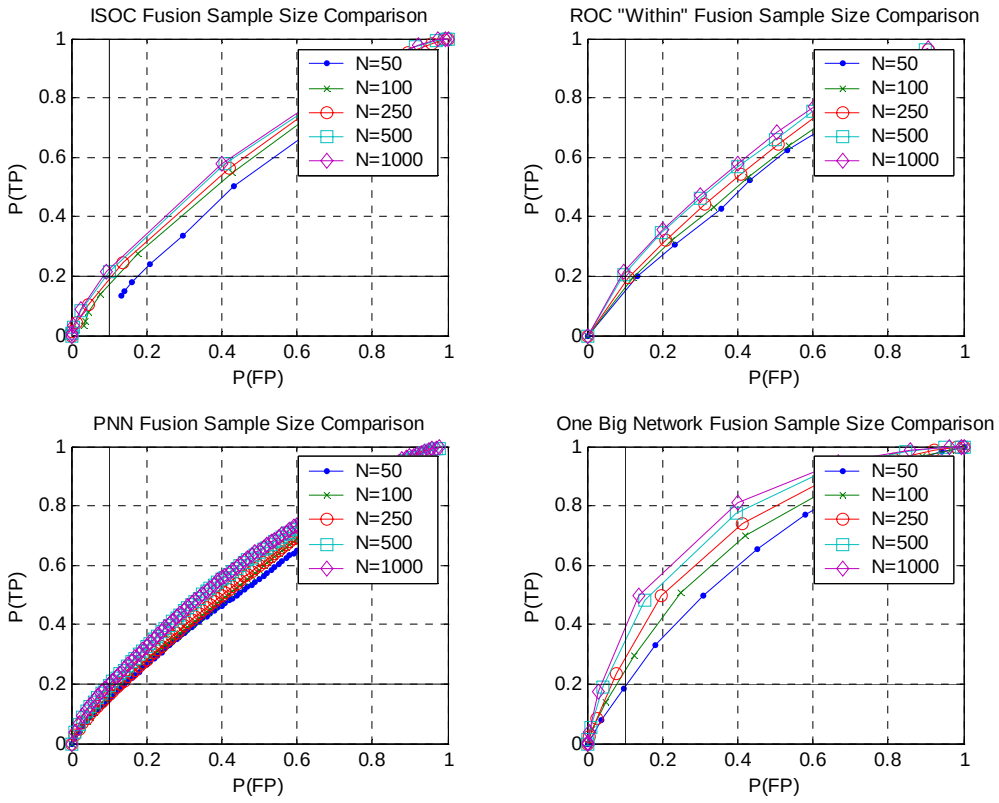


Figure 33: 8 Feature XOR ROC Curves, Across Sample Sizes.

As the problem becomes even more complicated, more samples are needed. For the ISOC, ROC "Within," and PNN Fusion Methods, there seems to be increasing performance between the first sample sizes. After a sample size of 500 in each class is reached, the performance stops increasing. The OBN shows a similar pattern. There is definitely a smaller return in performance from increasing the sample size from 500 to

1000 exemplars in each class, but it may be possible to increase performance a little more by further increasing the sample size.

Problem 6 Results: 8 Feature XOR with Autocorrelation Case, Single Sample Size

Features were generated according to the methodology described in the Data Generation: 8 feature XOR with autocorrelation, and the fusion process was followed as described in the Experimental Design section. Figure 34 shows four plots, one for each of the four fusion methods. They are the values of true positive for a false positive value of 0.1 on the average ROC curves over five replications with 1000 exemplars in each class. The true positive value is plotted against the level of across correlation. Each plot contains three lines, one for each of the three levels of autocorrelation.

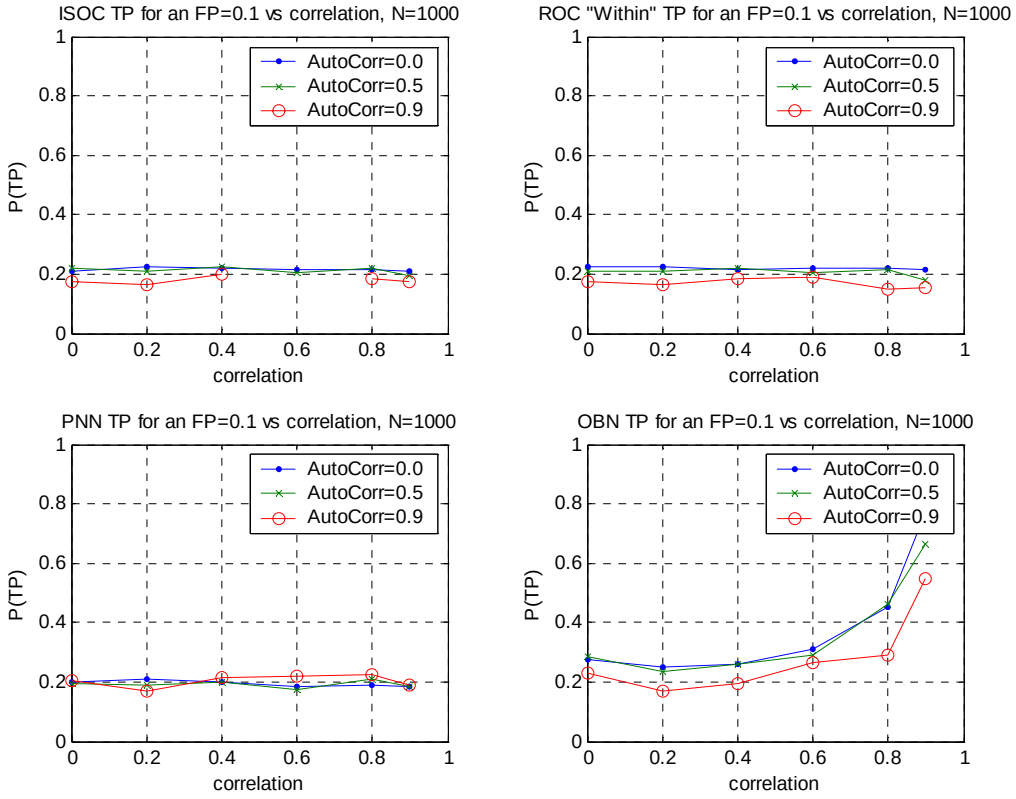


Figure 34: 8 Feature XOR with Autocorrelation Case, N=1000.

In this problem, as in previous problems, the ISOC Fusion and ROC "Within" are very robust to correlation. Also, as autocorrelation increases from 0.0 to 0.5, there is no degradation in performance for the ISOC and ROC "Within." As the autocorrelation increases from 0.5 to 0.9, there is a drop in performance across all levels of correlation, but it is a small decrease for both ISOC and ROC "Within." The PNN seems to be robust to both types of correlation. That is, all mostly flat and on top of one another. This is also the first problem that the PNN performance is as low as the ISOC and ROC "Within" performance. The OBN seems to be affected by both types of correlation, but it always performs as well or better than the other three methods at all combinations of the two types of correlation. As with the first two methods, it seems there is little difference

between an autocorrelation of 0.0 and 0.5, but there is a difference between an autocorrelation of 0.5 and 0.9. Also, there is an interesting pattern in performance for the OBN across levels of correlation for a given level of autocorrelation. The performance seems to decrease slightly at a correlation level of 0.2, but it increases as the geometry changes with the higher levels of correlation. The OBN continues to outperform the other methods, especially as the correlation level increases, in this type of problem because of the same geometric explanation in the previous problem. This is the first obvious example where it is better to eliminate the individual classifiers and treat the entire problem as One Big Network.

Problem 6 Results: 8 Feature with Autocorrelation Case, Across Sample Sizes

Features were generated according to the methodology described in the Data Generation: 8 feature XOR with autocorrelation, and the fusion process was followed as described in the Experimental Design section. This process was repeated for multiple sample sizes. Figure 35 shows four plots, one for each of the four fusion methods, They are the true positive values for a false positive value of 0.1 on the average ROC curves over the six levels of correlation for a given sample size. Sample size is varied from 50 exemplars in each class to 1000 exemplars in each class. Each plot contains three curves, one for each of the three levels of autocorrelation.

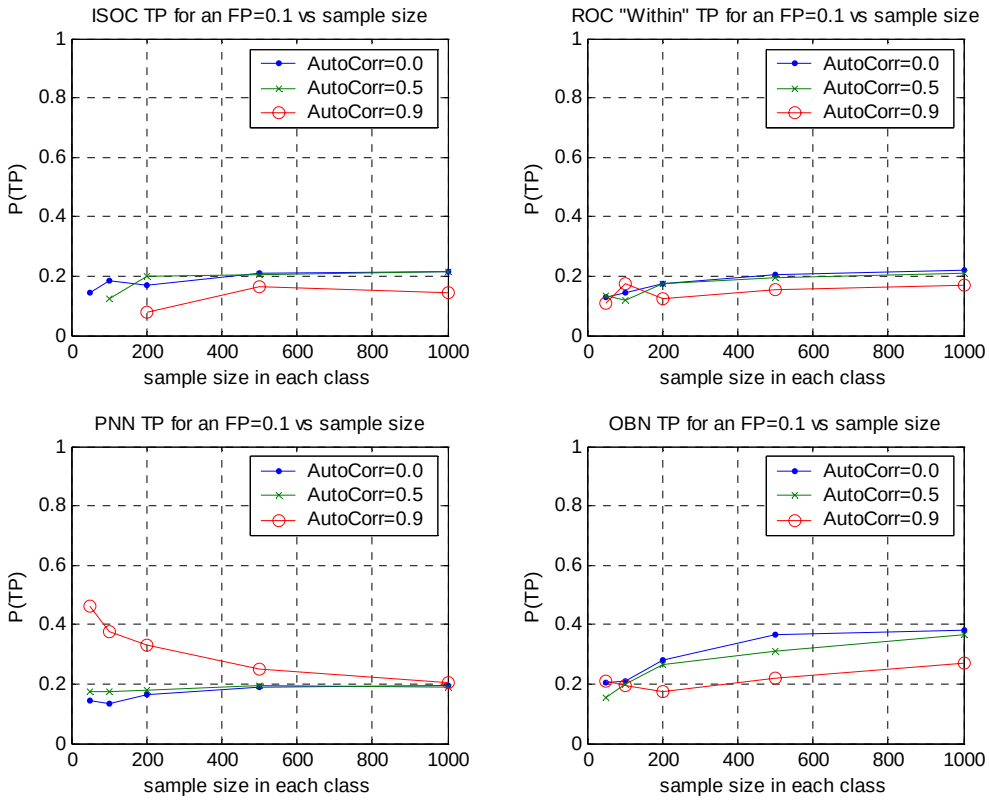


Figure 35: 8 Feature XOR with Autocorrelation Case, Across Sample Sizes.

For ISOC, ROC "Within," and OBN fusion, the performance decreases as the sample size decreases. While this degradation in performance is not dramatic, it is still present. There are a few exceptions, but this is mostly true across all levels of autocorrelation for these three methods. The PNN shows a much different result. For the two lower levels of autocorrelation, it follows the same trend the other three methods; as the sample size decreases, the performance decreases. There is an anomaly for the high level of autocorrelation that is explained by further examination of the posterior probabilities. Just as when the posterior probabilities where correlated with approximately the same level of across correlation as was inputted, the posterior probabilities are also approximately autocorrelated as the level of autocorrelation

inputted. Figure 36 shows the feature space of the PNN fusion from a single run at 0.0 level of autocorrelation, and Figure 37 shows the feature space of the PNN fusion from a single run at 0.9 level of autocorrelation.

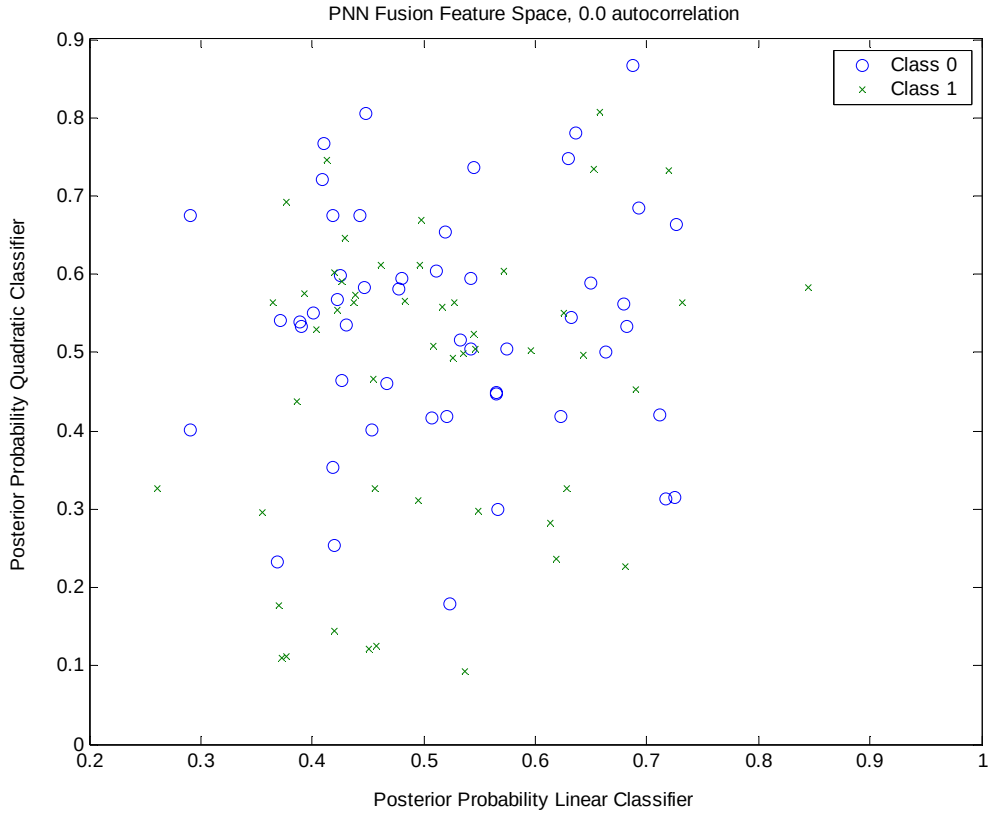


Figure 36: PNN Fusion Feature Space, 0.0 Autocorrelation.

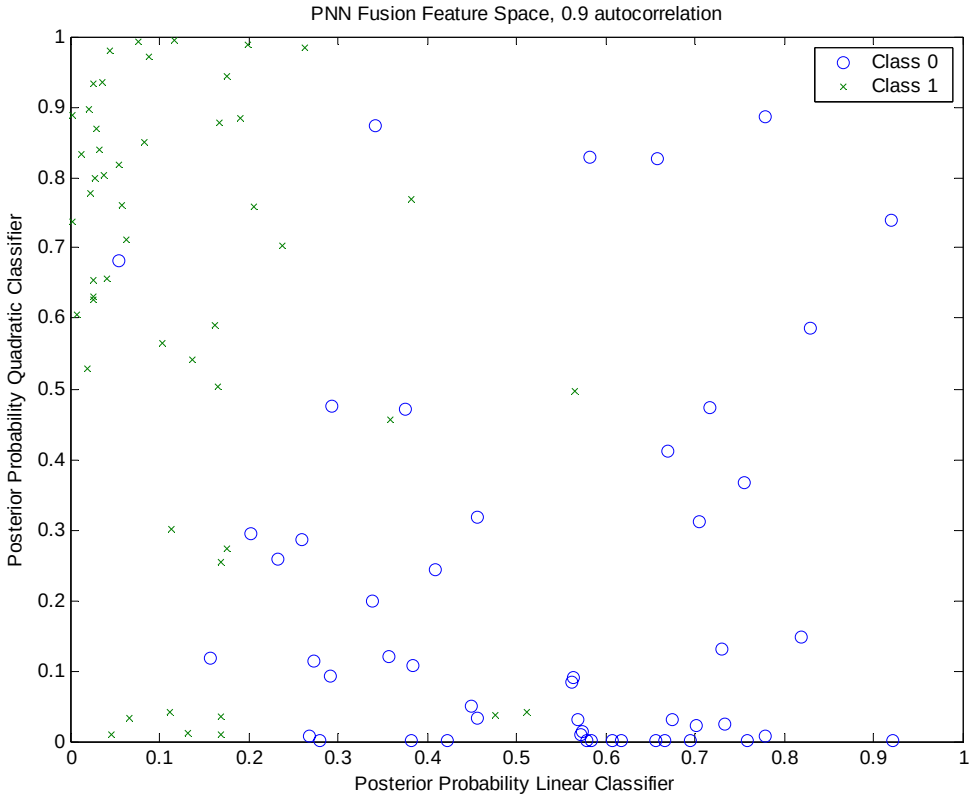


Figure 37: PNN Fusion Feature Space, 0.9 Autocorrelation.

Since the posterior probabilities are not autocorrelated in Figure 36, there is little separation between the classes; however, since the posterior probabilities are highly autocorrelated in Figure 37, there is a great deal of separation between the classes. In Figure 36 the autocorrelation of the posterior probabilities from the quadratic classifier, for instance, are -0.09, essentially 0. In Figure 37 the autocorrelation of the posterior probabilities from the quadratic classifier, for instance, are 0.83. This is especially evident at the low sample sizes because with only limited sample sizes at the high autocorrelation level, the features do not have enough observations to recover from the high autocorrelation levels. With the high samples, there will eventually be a great deal

of overlap between the two classes in the feature space. This explains the anomaly with the PNN fusion in this problem.

Problem 6 Results: 8 Feature XOR with Autocorrelation Case, An ANOVA Approach 

This same set of data can also be examined using an ANOVA approach. Consider a three factor design where the three factors are Level of Autocorrelation, Level of Across Correlation, and Sample Size. Level of Autocorrelation has three levels, 0.0, 0.5, and 0.9. Level of Across Correlation has six levels, 0.0, 0.2, 0.4, 0.6, 0.8, and 0.9. Sample Size has five levels, 50, 100, 250, 500, and 1000 exemplars in each class. A full factorial design consists of all possible combinations of these three factors. Each design point was replicated five times, and the response variable is the average response over the five replications. Table 8 summarizes the results of each ANOVA.

Table 8: Summary of XOR ANOVA Results, Averaged.

Method	**ISOC**	**ROC "Within"**	**PNN**	**OBN**
R^2	0.874706	0.768371	0.914785	0.944138
Adjusted R^2	0.713143	0.484626	0.810397	0.875708
Mean Response	0.149846	0.162378	0.223818	0.261231
Root MSE	0.027504	0.030867	0.042539	0.04626
Factors Considered and corresponding p-values				
Autocorrelation	<.0001	0.0009	<.0001	<.0001
Sample Size	<.0001	<.0001	0.0014	<.0001
Auto*Sample	0.0101	0.1110	<.0001	0.0004
Correlation	0.0183	0.0493	0.8690	<.0001
Auto*Corr	0.4385	0.8881	0.3466	0.0003
Corr*Sample	0.5225	0.7314	0.8049	<.0001

Again, in terms of significant variables, the results of the ANOVA confirm the results that were shown graphically. For the ISOC Fusion, autocorrelation, sample size,

and their two way interaction have the highest significance. That is, a change in these variables is most likely to trigger a change in the response. Although autocorrelation did not appear significant in the high sample size case shown graphically, it has a low p-value. Across correlation is the least significant of the main effects. For ROC "Within" Fusion, autocorrelation and sample size are the most significant which is what was shown graphically above. For PNN Fusion, autocorrelation, sample size, and their two way interaction have the highest significance. Again, this confirms the graphical results. For the OBN, all variables are significant. The OBN seems to be the most sensitive to changes in all three main effects. Overall, the OBN has the highest mean response, and the PNN has the second highest mean response. ISOC has the lowest mean response, and ROC "Within" has the second lowest mean response. The OBN has the highest R^2 values, and the PNN has the second highest R^2 values. The ROC "Within" has the lowest R^2 values, and the ISOC has the second lowest R^2 values. This different approach to looking at the same data a different way seems to confirm what was already show graphically.

Validation of the assumptions, normal errors and constant variance, via residual analysis was done for each of the four methods. All four followed the same pattern so only those for ISOC are shown. Figure 38 shows a histogram of the residuals; Figure 39 shows the residuals vs Row Number.

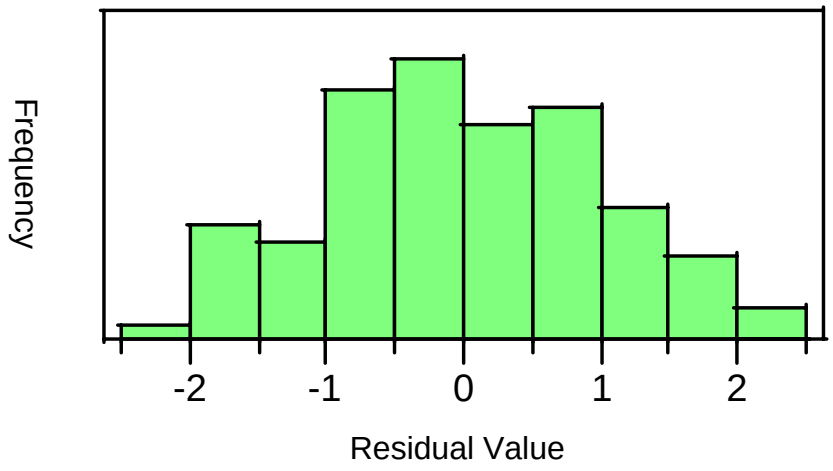


Figure 38: ISOC Histogram of Residuals.

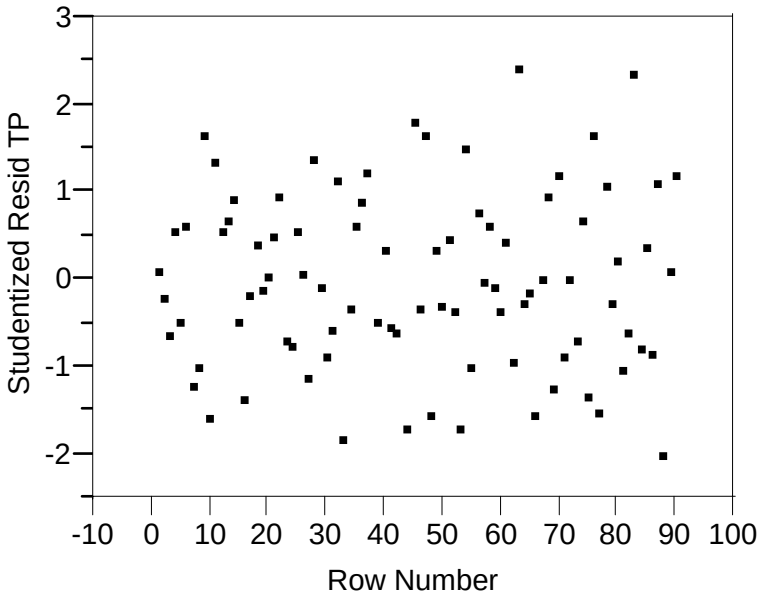


Figure 39: ISOC Residual TP Probability vs Row Number.

Figure 38 shows that the residuals are approximately normally distributed. Figure 39 shows that the residuals have approximately constant variance. These figures show that the two assumptions of the model, normal errors and constant variance, hold for this analysis.

Problem 7 Results: 20 Feature without Autocorrelation Case using Feature Selection

Features were generated according to the methodology described in the Data Generation: 20 feature without autocorrelation, and the fusion process and feature selection process was followed as described in the Experimental Design section. This process was repeated for two sample sizes over 15 replications. For each method and sample size, the average ROC curve was calculated with and without feature selection (twelve total ROC curves). Each plot contains six ROC curves, one for each level of correlation. Figure 40 shows the six ROC curves for sample size 50 in each class, and Figure 41 shows the six ROC curves for sample size of 1000 in each class.

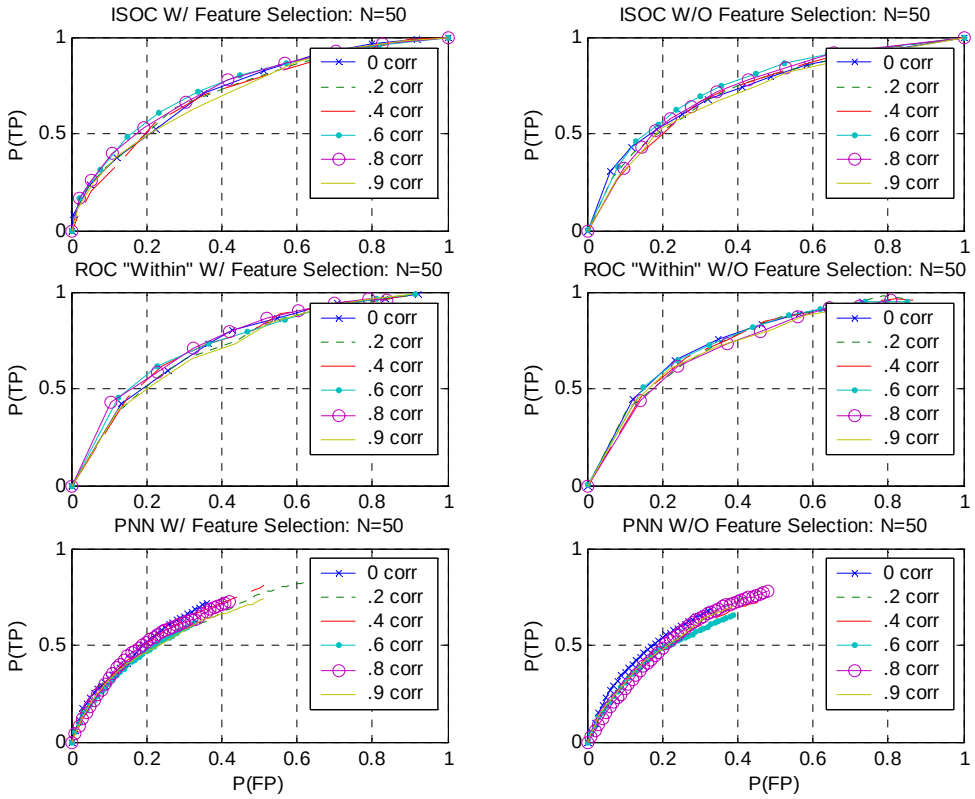


Figure 40: 20 Feature without Autocorrelation Case, N=50.

From these plots, it is obvious that in this particular problem, reducing the dimensionality of the feature set does not decrease performance in terms of ROC curves. There is not a big difference between the ROC curves with feature selection and without feature selection. This means that the feature set can be significantly reduced via feature selection without decreasing fusion performance.

For the low sample size problem, the feature selection process was not consistent. Sometimes the good features had high loadings, and sometimes they had low loadings. Sometimes the redundant features had high loadings, and sometimes they had low loadings. Sometimes even the noise features had high loadings, and sometimes they had low loadings. Regardless, similar performance was obtained using a significantly reduced feature set resulting from the feature selection.

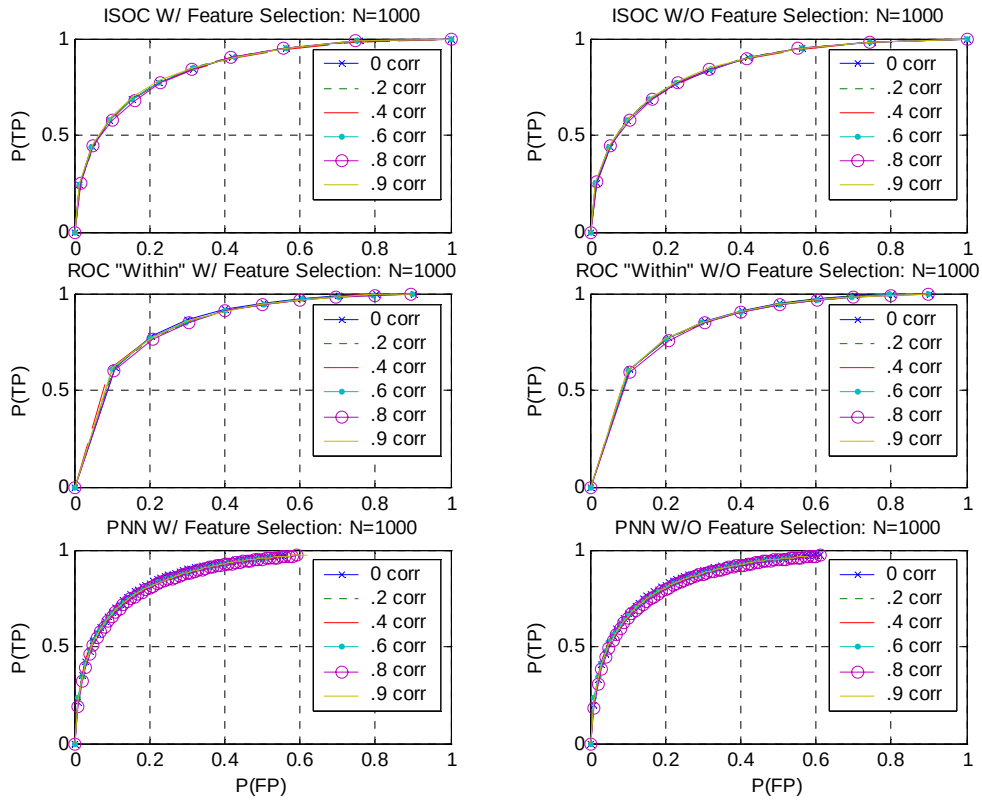


Figure 41: 20 Feature without Autocorrelation Case, N=1000.

From these plots, as in the low sample size problem, it is obvious that in this high sample size problem, reducing the dimensionality of the feature set does not decrease performance in terms of ROC curves. There is not a big difference between the ROC curves with feature selection and without feature selection. This means that the feature set can be significantly reduced via feature selection without decreasing fusion performance.

For the high sample size problem, the feature selection process was very consistent. In all fifteen runs, both the good features and all four redundant features had loadings greater than 0.45. This means that the feature selection process was able to

detect and delete all the noise features but not the redundant features. In addition, similar results were obtained using a much smaller feature set resulting from feature selection.

Since it is hard to visually compare the ROC curves, Figure 42 shows the value of true positive on the ROC curve for a false positive value of 0.1 for all six values of correlation for three methods for the low sample size problem. Figure 43 shows the value of true positive on the ROC curve for a false positive value of 0.1 for all six values of correlation for three methods for the high sample size problem.

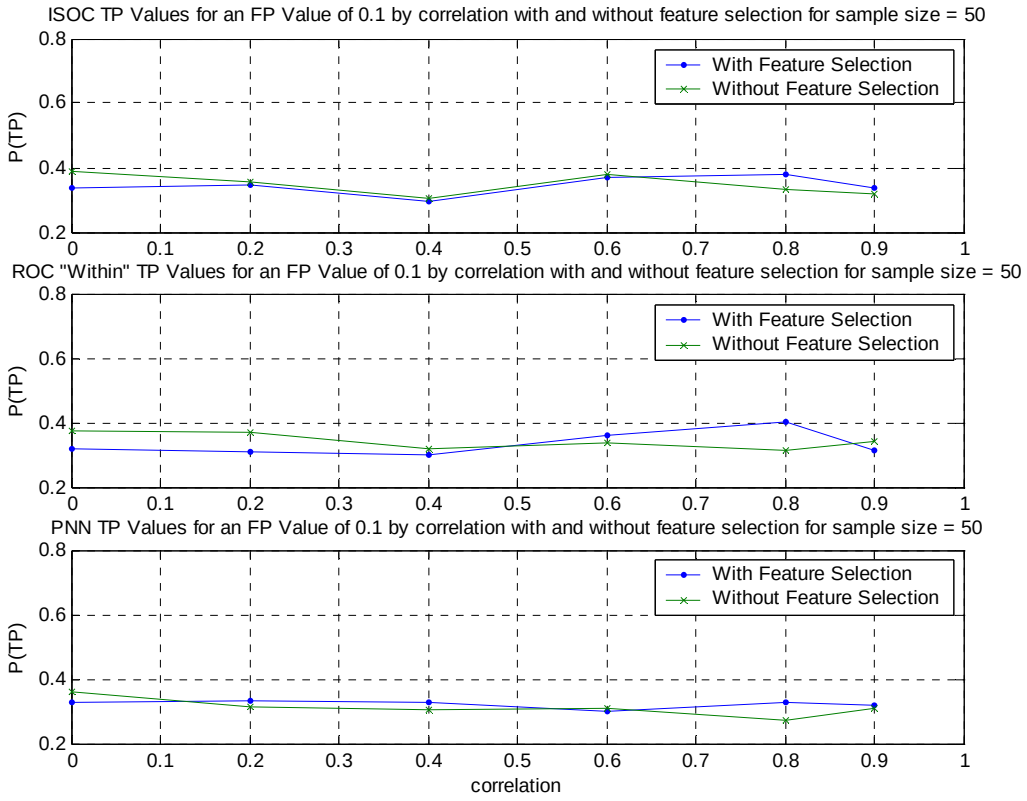


Figure 42: True Positive Values vs. Correlation Level for 0.1 False Positive Rate.

In Figure 42, it is apparent that reducing the feature set via feature selection results in little or no degradation in performance at the low sample size; however, regardless of feature selection, performance remains nearly the same across all levels of

correlation. In addition, this shows that these three fusion methods are fairly robust to correlation; that is, the value of true positive stays nearly constant across all levels of correlation. Also, the PNN performs comparably to the other two methods despite being given only half the data that the other two methods are given.

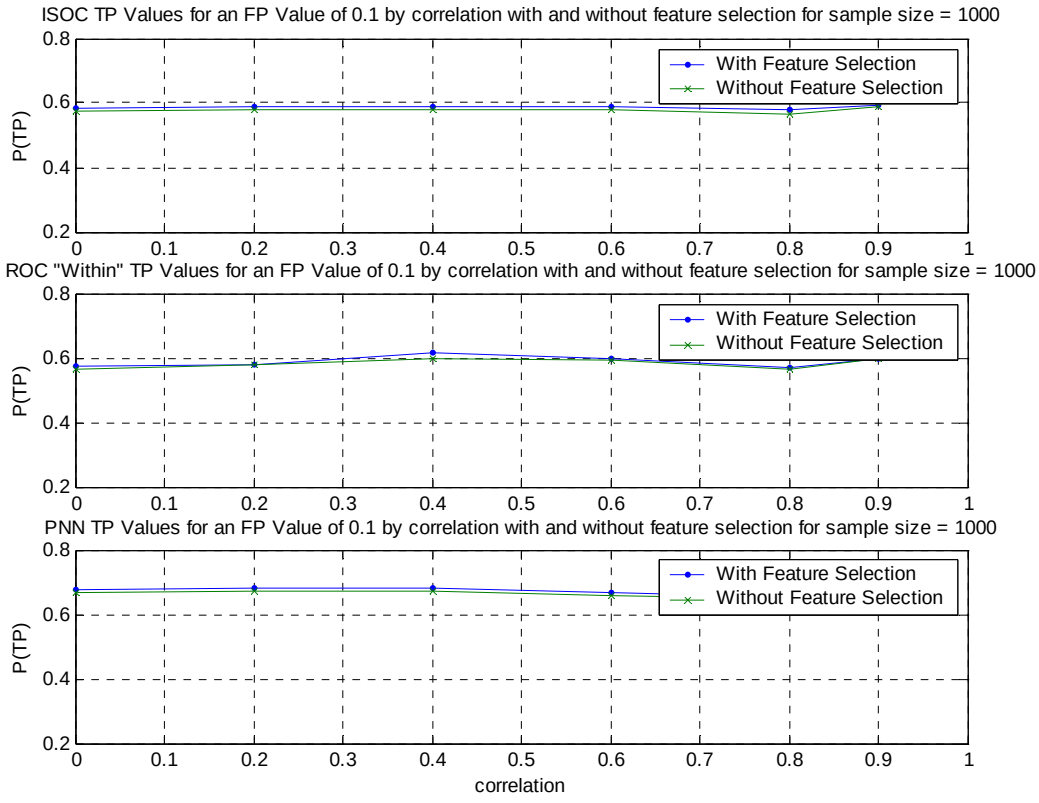


Figure 43: True Positive Values vs. Correlation Level for 0.1 False Positive Rate.

Figure 43 shows that, in the high sample size problem, performing feature selection always does as good or better than no feature selection; although, the improvement is minimal across all levels of correlation. In addition, this shows that these three methods are fairly robust to the level of correlation; that is, the value of true positive stays nearly constant across all levels of correlation. Also, the PNN outperforms

both the ISOC and ROC "Within" Fusion methods despite being given only half the data of those methods.

Problem 8 Results: 36 Feature without Autocorrelation Case, Single Correlation, Single Sample Size, using Feature Selection 

For this problem, the data was generated and the loadings were calculated as described above. First, the classification accuracy was calculated for each classifier fusion method using all 18 features for each classifier. Next, the features with the lowest six loadings were excluded and the classification accuracy was recalculated. Then, the classification accuracy was calculated for one feature up to twelve features for each individual classifier (e.g., when the number of features was five, the three methods used the five features with the highest loadings). These values were plotted for each fusion method in Figure 44.

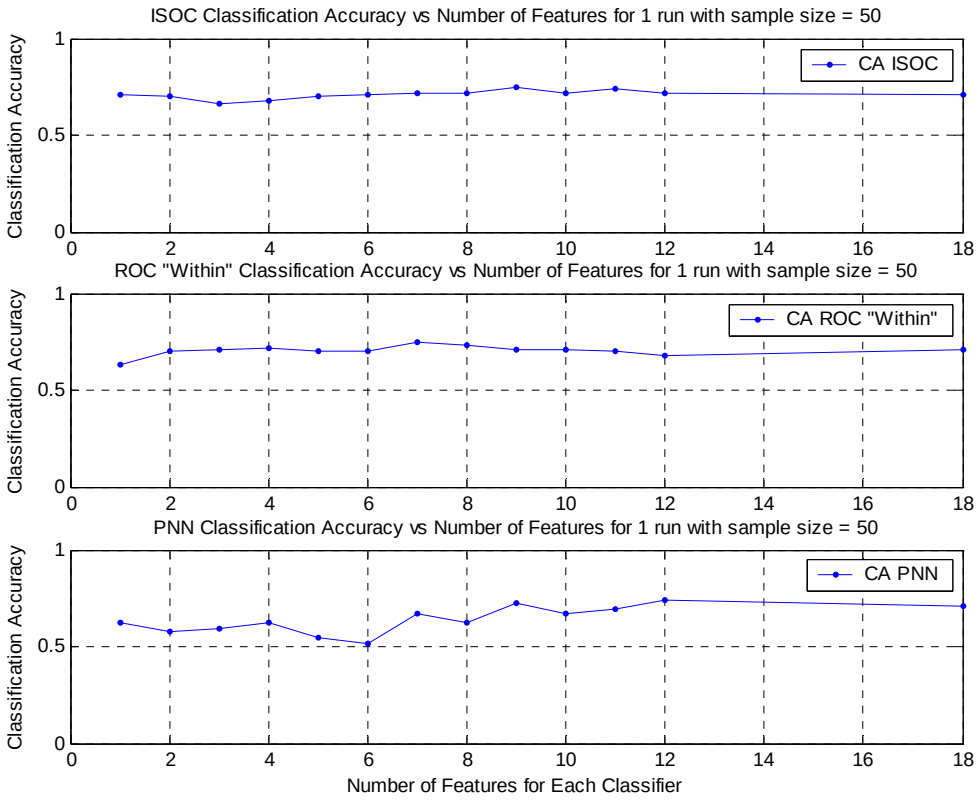


Figure 44: Classification Accuracy vs. Number of Features for 3 Fusion Methods.

In all three plots, the graph is relatively flat, indicating that in this run, the classification accuracy is relatively insensitive to the number of features that the individual classifiers observe. This means that, potentially, very few features could be used for the individual classifiers underlying the fusion, via feature selection, while maintaining the same level of performance as having many features. There seems to be some rising and falling of classification accuracy, but all increases and decreases are extremely mild.

Problem 9 Results: TPM Exploration 

An observation was made in all or part of the above analysis. There seems to be a declining trend in terms of fusion performance (i.e., ROC curves) as the correlation

between feature sets increases, but there also seems to be a point, usually at a higher level of correlation, where fusion performance actually benefits from the level of correlation between feature sets. To more fully understand this phenomenon, the TPM was calculated for different values of correlation between features for two of the problems already explored above. Figure 45 shows the TPM values vs correlation values for the 4 feature without autocorrelation case.

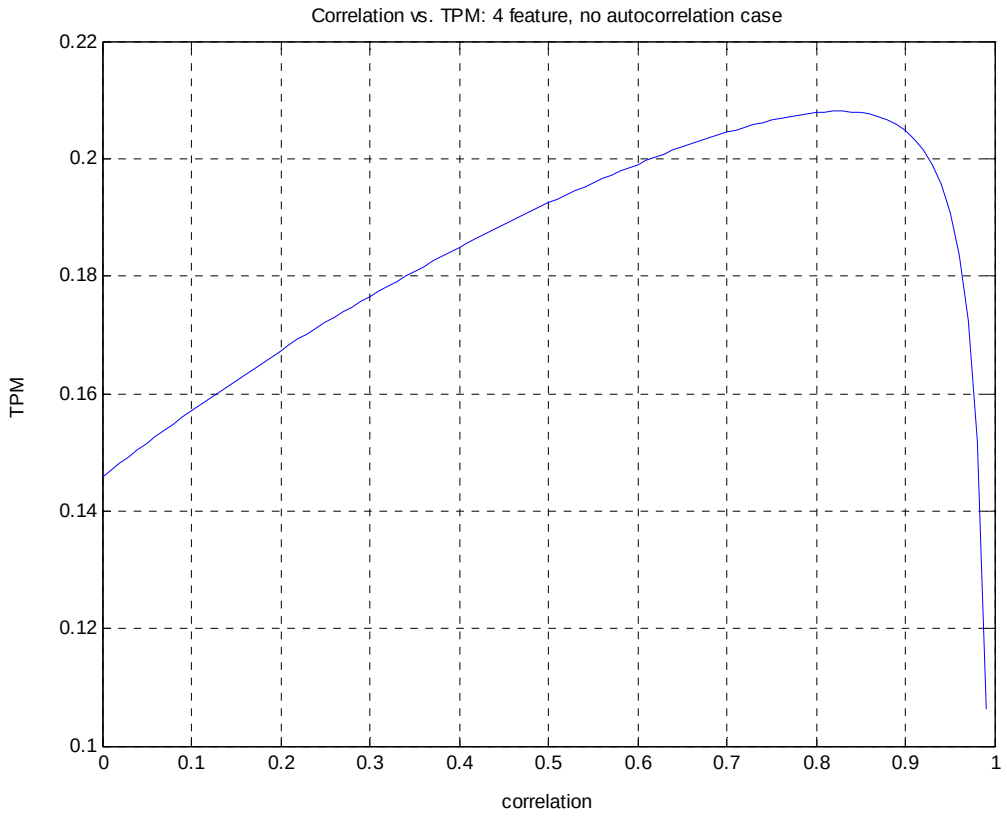


Figure 45: TPM Values vs Correlation, 4 Feature No Autocorrelation Case.

In Figure 45, it is apparent that the TPM rises as the correlation increases from 0.0 to 0.84, but there is a sharp decline after this point. The TPM at 0.99 correlation actually drops below the TPM at 0.00 correlation.

This led to additional experiments where the correlation between feature sets had only three values: one at 0.00 correlation, one at the highest point of the TPM plot, and one at 0.99 correlation. This process was replicated 5 times, and the average ROC curves were calculated. This was done at two sample sizes: 1000 exemplars in each class and 50 exemplars in each class. Figure 46 shows the average ROC curve for each fusion method with 1000 exemplars in each class for the 4 feature problem. Figure 47 shows the average ROC curve for each fusion method with 50 exemplars in each class for the 4 feature problem.

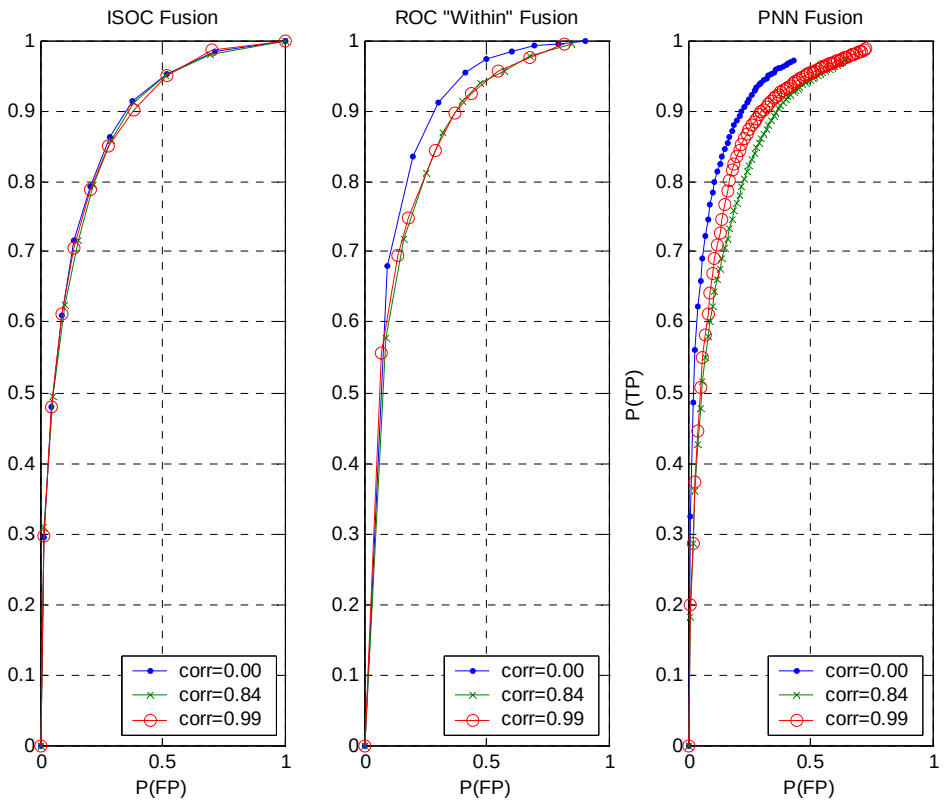


Figure 46: 4 Feature Problem, N=1000.

The ISOC plot from Figure 46 shows again that the ISOC Fusion is the most robust to correlation. It is very hard to tell the difference between any of the three

correlation levels. Thus, the ISOC plot does not show the results expected from the TPM calculations. The ROC "Within" plot shows that the 0.00 correlation curve is better on average than the two higher correlation levels, but the two curves for 0.84 and 0.99 correlation levels are almost identical. Thus, the ROC "Within" plot does not show the results expected from the TPM calculations. The PNN plot finally shows some separation between the three correlation levels. The 0.99 correlation was expected to outperform both the 0.00 correlation and the 0.84 correlation. While the 0.99 correlation does not outperform the 0.00 correlation, it does outperform the 0.84 correlation.

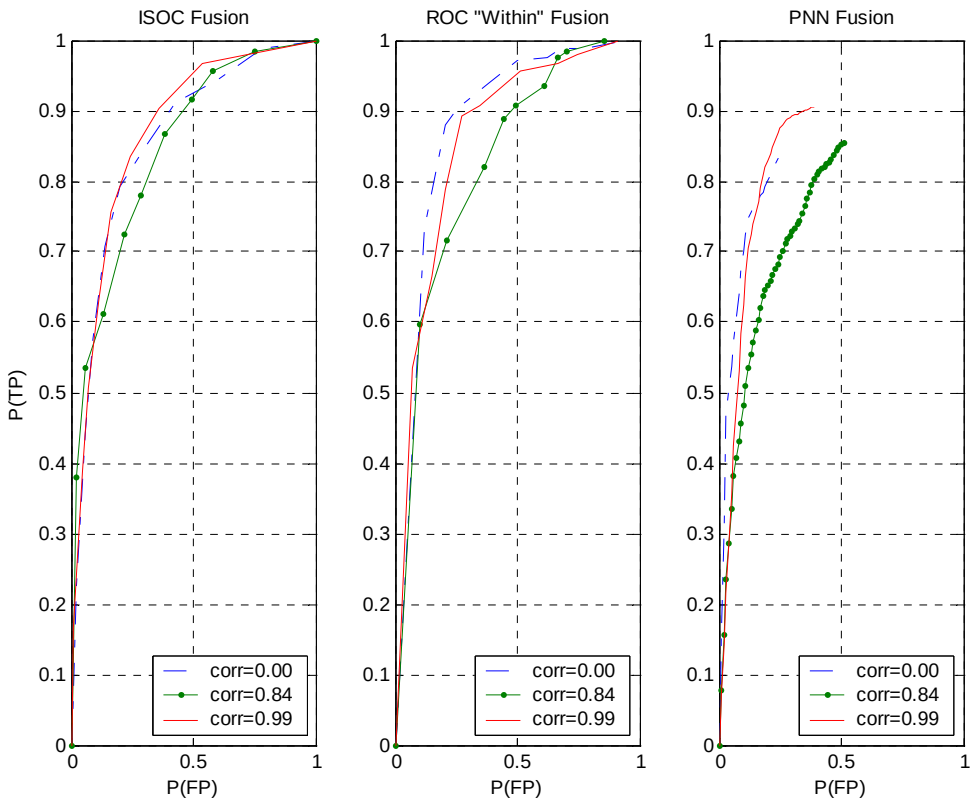


Figure 47: 4 Feature Problem, N=50.

The ISOC plot from Figure 47 shows a little different result than Figure 46. The 0.99 correlation actually outperforms the other two correlations at some points on the

ROC curves. This is what was expected from the TPM calculations. The ROC "Within" plot shows nearly the same results as Figure 46. While there is more separation between the higher correlation levels, the 0.00 correlation still outperforms the 0.99 correlation. The PNN plot in Figure 47 also shows nearly the same result as Figure 46. The 0.99 correlation outperforms the 0.84 correlation, but it does not outperform the 0.00 correlation.

Chapter Summary

This chapter provided the details of the findings and analysis of this thesis. Results from each of the problems posed in Chapter 3 were presented in this chapter. Insights resulting from each analysis were also given.

V. Conclusion

Introduction

This chapter concludes the thesis research. First, the major literature review findings are presented, and the general methodology employed is reviewed. The major results of this research are summarized, and recommendations for future research are discussed.

Literature Review Findings

In the current day, the United States Air Force is focused on accurate and timely targeting. Air Force Doctrine states that targets should not be struck with only single source intelligence (AFPAM 14-210, 1998). Instead, intelligence information from more than one source should be fused together in order to ensure a higher degree of accuracy (AFPAM 14-210, 1998). This higher degree of accuracy ensures less fratricide in combat operations.

There are many different ways to fuse data, and data can be fused at a variety of different levels. Many of these methods of fusing data assume that the inputs to the fusion are independent, and little is known about what happens when the inputs to the fusion are not independent (Willett, et al, 2000). In this research, four different ways of fusing information are exercised. The first two models, ISOC fusion (Haspert, 2000) and ROC "Within" fusion (Oxley and Bauer, 2002) are classifier fusion techniques that assume that the individual classifiers are independent. The third fusion method, PNN fusion, is a classifier fusion technique that makes no assumption about the independence of the classifiers. The fourth fusion method, OBN fusion, is not a classifier fusion

technique. It simply treats all the individual features as inputs to one big network. It makes no assumption about the independence of the features.

Methodology Employed

Data was generated for a variety of problems for this thesis. Correlation was introduced in two forms to the process: across correlation and autocorrelation. The level of these correlations were varied to observe how each method reacted to the correlation and to observe how these methods compared to each other at the same levels of correlation. In addition, sample size was varied throughout. In some cases, a feature selection process was performed to compare how the fusion performed in the presence and absence of feature selection both within and across the fusion methods. Finally, some explorations in TPM were performed.

Results

This thesis yielded many interesting results and a great deal of insight into the fusion process was obtained. Problem 1 and problem 2 possess a very similar structure, and they both provide similar insight. They both show that, in this type of problem, the PNN and OBN are superior to the ISOC and ROC "Within" fusion methods at all levels of across correlation. While the ISOC and ROC "Within" fusion methods are very robust to the across correlation, they do not perform as well as the other two methods. In fact, the PNN outperforms the other three methods while only observing half the data as the other three methods. This trend is true regardless of sample size.

Problems 3 introduces autocorrelation into the fusion process. Despite this addition of autocorrelation, all four methods observe the same trends in terms of across correlation and sample size as those observed in problem 1 and problem 2. In problem 3,

ISOC, ROC "Within," and OBN are susceptible to autocorrelation, especially at low sample sizes. The PNN seems to be robust to autocorrelation in this type of problem.

Problem 4 is the first problem where something unexpected occurred. For all four methods, increasing the level of correlation actually improves the performance of the fusion. This is easily explained with a geometric interpretation of the problem. Again, the ISOC and ROC are very robust to the across correlation, but they are always outperformed by the PNN and OBN. This is true across all sample sizes.

Up until this point, the PNN and OBN had been performing very similarly; they always outperformed the other two methods. In problem 5, the PNN only performs as well as the ISOC and ROC "Within" fusion methods; on the other hand, the OBN outperforms the other three methods at all levels of correlation. Problem 6 showed that the OBN continued to outperform the other three methods in the presence of autocorrelation. This is true in almost every case; the PNN outperforms the OBN at low sample size cases with high autocorrelation. This is another case that is counter-intuitive, but it is easily explained with a geometric interpretation.

Problem 7 and problem 8 are two problems in which the number of features is increased so that feature selection can be explored. In both problems, it was shown that feature selection can be very beneficial. Feature selection was used to reduce the dimensionality of the problems without degradation in fusion performance.

Problem 9 is a further investigation into problem 1 at only specific levels of across correlation. There are cases where the TPM will actually decrease at very high levels of across correlation. The results from problem 9 show that the fusion for a highly

correlation feature set can actually be better than the fusion for a more moderately correlated feature set.

Overall, some key insights were gained from this research. First, the across correlation injected into the fusion process is approximately equal to the correlation of the posterior probabilities of the individual classifiers. Second, the autocorrelation injected into the fusion process is approximately equal to the autocorrelation of the posterior probabilities of the individual classifier. Next, while a high level of correlation is usually associated with having less information, sometimes this high level of correlation actually aids the fusion as it changes the geometry of the problem. Next, ISOC and ROC "Within" seem to be the most robust to across correlation even though they are the methods that assume independence of the classifiers. Although, they are robust methods, they are always outperformed by one of the other two methods which do not make the independence assumption. High levels of autocorrelation seemed to decrease performance of each of the fusion methods for all sample sizes except the PNN at low sample sizes. Also, generally a lower sample size results in lower performance, and usually, there is a point where adding more samples will not necessarily increase performance. Finally, OBN seems to be the most successful fusion method as it performed as well or better than the other three methods for each of the problems in this thesis.

Recommendations for Future Research

While this thesis provided a great deal of insight into the fusion process, there is still much more research that can be done in the field. First, the biggest shortcoming of this research is that all the data was fabricated. As real-world data sets become available,

this fusion process should be applied to those data sets. In the absence of real world data, the feature sets used in this thesis could be expanded so that more noise and redundant features are added. Next, different classifiers, such as neural networks, could be used as the individual classifiers instead of using the linear and quadratic discriminant functions. Also, in concurrent thesis research, fusion has been done with three classifiers, but this could be extended even further to a larger number of classifiers. Once the number of classifiers has been extended, classifier selection, similar to feature selection, can be performed to select only good classifiers to be fused. Finally, all of this research focuses on the two-class problem; research could be extended to a three-class or higher problem.

Bibliography

Air Force Doctrine Document 2-1, *Air Warfare*, 22 January 2000.

Air Force Pamphlet 14-210, *USAF Intelligence Targeting Guide*, 1 February 1998.

Haspert, J.K., "Optimum ID Sensor Fusion for Multiple Target Types." *IDA Document D2451*, 2000.

Clutz, T., *A Framework for Prognostics Reasoning*. Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 2002.

Fukunaga, Keinosuke, and Raymond R. Hayes, "Effects of Sample Size in Classifier Design." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Volume II, No. 8, August 1989, pp. 873-885

Laine, Trevor, "Relevant Background for the Research of Sensor Fusion as Applied to Automatic Target Detection/Recognition (ATD/R) in an Environment with Correlated Input Data." *Advanced Applications for ANNs Special Study,* 2003.

Oxley, M.E., and K. Bauer, "Classifier Fusion for Improved System Performance." *AFIT/ENS Working Document 02-02*, 2002.

Ralston, J.M., "Bayesian Sensor Fusion for Minimum-Cost I.D. Declaration." *1998 Joint Service Combat Identification Systems Conference on Requirements, Technologies and Developments*, (CISC-98), Volume 1 – Technical Proceedings.

Shipp, C.A and L.I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers." *Information Fusion*, vol 3, iss 2: 2 June 2002, pp. 135-138.

Storm, S. A., K. Bauer, and M. Oxley, "An Investigation in the Effects of Correlation in Classifier Fusion." *Intelligent Engineering Systems Through Artificial Neural Networks,* vol 13: 2003, pp.619-624.

Wasserman, P., and V. Nostrand, *Advanced Methods in Neural Computing.* 1993.

Willet, P., P.F. Swaszek, and R.S. Blud, "The good, bad and ugly: distributed detection of a known signal in dependent Gaussian noise." *IEEE Transactions on Signal Processing*, vol 48, iss 12: 12 December 2000, pp. 3266-3279.

Vita

Captain Nathan J. Leap graduated from Bishop McCort High School in Johnstown, Pennsylvania. He entered undergraduate studies at the United States Air Force Academy (USAFA) where he graduated with a Bachelor of Science Degree in Operations Research in June 1999. He was commissioned through USAFA with a Reserve Commission.

His first assignment was at Eglin AFB as an operations research analyst for the 28th Test Squadron, 53rd Test and Evaluation Group, 53rd Wing. In August 2003, he entered the Graduate School of Engineering and Management, Air Force Institute of Technology. Upon graduation, he will be assigned to Air Force Materiel Command, Wright-Patterson AFB, Ohio.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
03-2004	**Master's Thesis**	Jun 2003 – Mar 2004

4. TITLE AND SUBTITLE

AN INVESTIGATION OF THE EFFECTS OF CORRELATION, AUTOCORRELATION, AND SAMPLE SIZE IN CLASSIFIER FUSION

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Leap, Nathan J., Capt., USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 641
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GOR/ENS/04-06

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR
Attn: Major Juan R. Vasquez
801 North Randolph Street, Room 933 (703) 696-8431
Arlington, VA 22203-1977 e-mail: Juan.Vasquez@afosr.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This thesis extends the research found in Storm, Bauer, and Oxley, 2003. Data correlation effects and sample size effects on three classifier fusion techniques and one data fusion technique were investigated. Identification System Operating Characteristic Fusion (Haspert, 2000), the Receiver Operating Characteristic "Within" Fusion method (Oxley and Bauer, 2002), and a Probabilistic Neural Network were the three classifier fusion techniques; a Generalized Regression Neural Network was the data fusion technique. Correlation was injected into the data set both within a feature set (autocorrelation) and across feature sets for a variety of classification problems, and sample size was varied throughout. Total Probability of Misclassification (TPM) was calculated for some problems to show the effect of correlation on TPM. Feature selection was performed in some experiments to show the effects of selecting only certain features. Finally, experiments were designed and analyzed using analysis of variance to identify what factors had the most significant impact on fusion algorithm performance.

15. SUBJECT TERMS

Sensor Fusion, Classifier Fusion, Classification, Probabilistic Neural Network, ISOC Fusion, ROC Curve Fusion, Correlation, Autocorrelation, Sample Size

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Kenneth W. Bauer, AFIT/ENS
U	U	U	UU	119	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-6565, ext 4328; e-mail: Kenneth.Bauer @afit.edu